The Use of Pseudo Panel Data for Forecasting Car Ownership

Huang, Biao
Department of Economics, Birkbeck College, University of London

June 2007

The Use of Pseudo Panel Data for Forecasting Car Ownership

Biao Huang

Birkbeck College, University of London

Thesis Submitted in Accordance with the Requirements for The Degree of Doctor of Philosophy

June 2007

Declaration

I confirm that the work presented in this thesis is my own and appropriate references have been made to the work of others.

Acknowledgement

I would like express my extreme gratitude to my supervisor, Prof. Ron Smith, whose insight and constructive comments have proved invaluable. Without his guidance and supervision, the completion of this thesis would be impossible.

The early research findings were presented to the European Transport Conference in 2005. The comments from various seminar participants as well as Dr. Walter Becket of Birkbeck College are gratefully received.

I also thank Dr. Gerard Whelan for generously supplying a copy of his PhD thesis. This work also benefits indirectly from my other projects work with colleagues at the Transport Studies Unit, University of Oxford, MVA Consultancy and Mr. John Bates. I thank them for their unknowing contributions.

Finally, I would like to thank my wife Dan Li for her enormous support during the whole period of my research.

Abstract

While car ownership forecasting has always been a lively area of research, traditionally it was dominated by static models. To utilize the rich and readily available repeated cross sectional data sources and avoid the need for scarce and expensive panel data, this study adopts pseudo panel methods. A pseudo panel dataset is constructed using the Family Expenditure Survey between 1982 and 2000 and a range of econometric models are estimated. The methodological issues associated with the properties of various pseudo panel estimators are also discussed.

For linear pseudo panel models, the methodological issues include: the relationship between the pseudo panel estimator and instrumental variable estimator based on individual survey data; the problem of measurement errors (and when they can be ignored) and the consistent estimation of dynamic pseudo panel parameters under different asymptotics. Static and dynamic models of car ownership are estimated and a systematic specification search is carried out to determine the model with best fit. The robustness of the estimator is investigated using parametric bootstrap techniques.

As an individual household's car ownership choice is discrete, limiting the model to linear form is obvious insufficient. This study attempts to combine the pseudo panel approach with discrete choice model, which has the distinctive advantages of allowing both dynamics and saturation but without the need for expensive genuine panel data. This does not seem to have been done before. Under the framework of random utility model (RUM), it is shown that the utility function of the pseudo panel model is a direct transformation from that of cross-sectional model and both share similar probability model albeit with different scale. This study also explores the various forms of true state dependence in the dynamic models and tackles the difficult econometric issues caused by the inclusion of lagged dependent variables. The pseudo panel random utility model is then applied to car ownership modeling, which is subsequently extended to take saturation into account. The model with the best fit has a Dogit structure, which is consistent with the RUM theory and is able to estimate the level of saturation and test its statistical significance.

Both linear and discrete choice models are applied to generate forecasts of car ownership in Great Britain to year 2021. While the forecasts based on discrete choice models closely match the observed car stock between 2001 and 2006, those based on linear models appear to be too high. Furthermore, the results from nonlinear models are comparable to the findings in other authoritative studies, while the long term forecasts from linear models are significantly higher. These results highlight the importance of saturation, and hence the choice of model functional form, in car ownership forecasts. In conclusion, we make some comments about the usefulness of pseudo panel models.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

The car market is an important sector of most modern economies and both the demand for ownership and the demand for new cars play an important role in economic decisions. The health of the car industry in general depends on consumers' demand for new cars. Demand forecasting is one of the most important tools car manufacturers use in their financial planning and decision making about expansions and contractions of plant capacity. For various government and public bodies, understanding and forecasting demand for car ownership are equally important. As the most important user of petroleum fuel, the car market has a strong influence on non-replaceable energy. Projections of future fuel consumption, and the impacts on fuel consumption of various forms of government intervention, are routinely based on forecasts for car ownership demand. Furthermore, understanding the factors driving demand for cars is important in addressing a range of environmental issues including local air pollution and climate change. Since car emissions are a large component of pollution, air quality standards and policies are largely based upon projected car ownership and use. Finally, accurate car demand models are also an aid to planners who must anticipate infrastructure needs, address concern of congestion and provide public transport services. Government agencies and local passenger transport authorities utilize projections of car ownership levels as a key input to obtain accurate projections of infrastructure needs and public transport patronage.

Car ownership forecasting plays a central role in the planning and decision making of numerous public agencies and private organizations. Given the important role of car demand forecasts in a wide variety of settings, it is not surprising that it has been a lively area of research and numerous models have been constructed to forecast car demand. It is important to recognize that the choices of model structure and functional form are heavily influenced by the objective and context of the study, and there is no single model that would offer best performance in all situations. For example, for short term forecast, a simple time series model based on aggregate data might perform better than the more complex disaggregate model. Also, while the consideration of saturation might not add much value for car forecasting in developing countries, it might be

highly significant in mature markets. It is easier to understand these points by looking at the car market in Great Britain as an example. Between 1950 and 2005, the total number of cars in the stock increased from 1.98 million to 26.21 million, implying an average growth rate of 4.7% per annum; for the same period, the real Gross National Income (GNI) increased from £243 billion to £987 billion (1995 prices), implying an average growth rate of 2.5% per annum (DfT, 2006c; ONS, 2007). If one uses certain time series models (for example, a simple Error Correction Model) for long term forecasts, it would substantially over-estimate the car stock in distant future. This is because the past growth trend will be inevitably curtailed by the approach of saturation: in 1951, only 14% of households had regular access to at least one car, while this proportion increased to 75% in 2004 (DfT, 2006c).

In the current study, the car demand forecasting model is developed within the context of British car market. In the UK, the Department for Transport has commissioned a number of "official" forecasting models over the past few decades, which include those developed by the Transport and Road Research Laboratory, the Regional Highway Traffic Model (RHTM), the National Road Traffic Forecasts (NRTF) and the National Transport Model (NTM). Besides the Department for Transport, there are industrial organisation such as SMMT (The Society of Motor Manufacturers and Traders) and other commercial and academic organisations, which are also involved in this area of research. However, some of the research remains "in house", i.e., the details of the model are not publicly available. Some of the studies available in the academic journals used methodology and data different to the NRTF/NTM, yet each study has its own limitations. Various types of car demand forecasting models developed in Great Britain and worldwide will be reviewed in Chapter 2.

The literature review reveals that the static approach dominates car ownership forecasting in Britain. The motivation for this thesis is that the inclusion of dynamics will yield fruitful results and lead to more accurate forecasts. Traditionally, empirical models of individual travel choice behaviour have been built on the assumption of equilibrium and suffered from a lack of dynamics. In the past two decades, the importance of dynamics in transport is gradually gaining recognition. In various areas of transportation research, issues such as the temporal dependence of choices, the role

of habit, imperfect information regarding alternatives and prices, costs of adjustment and transaction costs, have been empirically assessed.

Nevertheless, the use of dynamic approach in car demand forecasting is still limited due to heavy data requirements. Due to data constraint, there have been relatively few forecasting models that use the dynamic approach except those using aggregate time series methods. It is possible to forecast car demand using panel data models. However, there is only one panel survey in Britain containing limited transport related information: the British Household Panel Survey (BHPS), which is inadequate for the purpose of our study. Furthermore, due to the attrition problem, the size and representativeness of the samples decline over time, rendering the panel data inferior to other national cross-sectional data. For example, less than half of the respondents in Wave 1 of BHPS remained in the sample in Wave 13, and various population groups such as the old, the young, the unemployed, those with low income, etc. became significantly under-represented (ISER, 2006).

One approach to circumvent the need for panel data is to construct pseudo panels from the cross sectional data. The pseudo-panel approach is a relatively new econometric approach to estimate dynamic demand models. A pseudo-panel is an artificial panel based on (cohort) averages of repeated cross-sections. The cohorts are defined based on time-invariant characteristics of the households and extra restrictions should be imposed on pseudo-panel data before one can treat it as genuine panel data. Using the cohort data over a number of periods, one could distinguish long run and short run effects while allowing for heterogeneity between the cohorts. In this way, one is able to overcome the deficiencies in both the static models and aggregate time series.

The application of pseudo panel car ownership model raises many interesting questions. For example, how do we define the cohort so the econometric model is identified and the measurement errors in variables can be minimized (ignored)? There is a question about the treatment of the dependent variable. Whereas in the original data car ownership is a discrete variable (zero, one, two, ..) in pseudo panel data it is a continuous variable, e.g. average number of cars per household or the proportions of households that own cars. There is a question about the treatment of transformations of the variables. Should one use the average of the transformed micro data or transform

the pseudo panel cohort data? In many cases, e.g. logarithmic transformations, the average of the transformed data is not defined, since the micro data contain zeros[1]. Is it possible to apply the microeconomic theory of utility maximization for individual decision makers and combine the pseudo panel model with the random utility model? What are the pros and cons of discrete choice (nonlinear) pseudo panel model and what's its relationship to standard random utility models? How can the nonlinear pseudo panel model be consistently estimated? And finally, what are the empirical appeals of the pseudo panel models and how well do they perform in car ownership forecasting?

To facilitate readers' understanding, we list the most common notations that are used throughout the thesis in Table 1-1. The following subscripts are consistently used: i denotes the individual household in the micro survey; c denotes the associated cohort of household i; and t denotes years.

Table 1-1 Common Notations

Notation	Description
A_{ct}	Average number of cars (automobiles) per household in cohort c in year t
P_{1+}	Probability of household owning at least one car
$P_{2+\|1+}$	Probability of household owning two or more cars conditional on owning at least one car
n_{ct}	Number of sample observations in cohort c in year t
N_c	Total population of cohort c (assumed to be constant in the theoretical model, i.e. no birth or death)
C	Total number of cohorts
α	Scalar: coefficient for the lagged dependent variable
β	K x 1 vector of coefficients for exogenous explanatory variables
λ	Unobserved cohort heterogeneity (fixed effect or random effect)

In the empirical work, the dependent variable is A_{ct} for all linear models. It is slightly more complicated for discrete choice models. We observed the proportion (*not* probability) of households owning at least one car in cohort c in year t, which is noted as r_{ct}^{1+}. Among the car owning households in cohort c, we also observed the proportion of those owning two or more cars, which is noted as $r_{ct}^{2+|1+}$. They are the dependent

[1] There are also theoretical considerations on the treatment of variable transformations. Again, it remains a question whether the method proposed in the standard linear pseudo panel econometric literature should be applied to discrete car ownership choices.

variables in two separate discrete choice pseudo panel models. The first order condition of the maximum likelihood function implies that the predicted probability (P_{1+} and $P_{2+|1+}$) equals the proportions (r_{ct}^{1+} and $r_{ct}^{2+|1+}$) in the sample under certain conditions (e.g. probability model is a multinomial logit model with alternative specific constant).

The thesis is organized as follows. Chapter 2 is the literature review of car ownership models. Chapter 3 describes the data used in the thesis including the construction of the pseudo panel dataset. Chapter 4 discusses the linear static fixed effect models, investigating the relationship between the pseudo panel estimator and the instrumental variable estimator based on individual survey data as well as the measurement error problems (and when they can be ignored). Chapter 5 discusses the consistent estimation of linear dynamic pseudo panel model under different asymptotics and the rank conditions for identification. For empirical models of car ownership, systematic specification search is carried out to investigate various issues such as appropriate explanatory variables, functional forms, problems of heteroskedasticity and autocorrelation, fixed or random effects and presence of heterogeneity.

Chapter 6 extends the pseudo panel approach to discrete choice model. The pros and cons of nonlinear pseudo panel model are discussed and a pseudo panel model that is consistent with random utility theory is presented. Chapter 7 investigates dynamic discrete choice model of pseudo panel. Models with different forms of (true) state dependence are compared and consistent estimation methods are proposed for the preferred first order Markov model. For the car ownership model, saturation is an important concept so the theoretical model has been extended and a pseudo panel Dogit model is presented. Empirical models of households with at least one car and those with two or more cars conditional on owning the first car are estimated separately.

Chapter 8 uses both the linear and nonlinear econometric models to generate car ownership forecasts for Great Britain between 2001 and 2021. The forecasting results are compared to the observed car stock between 2001 and 2006 as well as the forecasts in other authoritative studies. A number of scenario tests are carried out to examine the sensitivity of the forecasting models. Chapter 9 is a brief conclusion, where the usefulness of pseudo panel models is also considered.

Chapter 2 Review of Car Ownership Models

In the chapter, the car ownership models are reviewed. In an early literature review by Train (1986), the methodology found in the literature was divided into two categories: disaggregate and aggregate. The disaggregate method comprises of compensatory and non-compensatory model. In the compensatory models, the household is assumed to trade off characteristics in the sense that a high value of one characteristic can compensate for a low value of another characteristic (for example, the household would choose a car that is smaller than it wants if the price is sufficiently low). The compensatory models can be based on either real choice situations (Revealed Preference Studies) or hypothetical choice situations (Stated Preference Studies). In the non-compensatory models, the consumer is assumed to have an importance ranking of characteristics of the alternatives, and, for each characteristic, have some minimum acceptable level, called a "threshold". In the decision making process, the consumer eliminates from consideration all the alternatives that do not meet the minimum acceptable standard (threshold) for the characteristic which he considers most important. If more than one alternative remains after the initial elimination, then the consumer looks at his second ranked characteristic and eliminates any alternatives that are not above the threshold for this characteristic. This process continues until only one alternative remains; this is the alternative which the consumer chooses.

According to Train (1986), the aggregate method comprises of approximate demand equations and consistent demand equations. The true aggregate demand function for an area is the sum of the demand functions for all the individuals in the area. It is called "consistent demand function" since it is consistent with underlying demand at the individual level. Due to the complexity of such functions, many aggregate models are derived by specifying an aggregate demand function that is not necessarily consistent with realistic individual demand equations and considering it an approximation to the true aggregate demand functions. In Train's review, two studies (Cardell and Dunbar, 1980; Boyd and Mellman, 1980) used consistent aggregate demand equations and estimated aggregate demand equations with explicit account taken of the fact that aggregate demand is the sums of individual demands.

Rand (2002) and De Jong et al (2004) classified the car ownership model into nine different types: Aggregate time series model, cohort model, aggregate market model, heuristic simulation model, indirect utility model, static disaggregate choice model, panel model, pseudo panel model and dynamic transaction model. These different model types are then compared based on a number of criteria: inclusion of demand and supply side of the car market, level of aggregation, dynamic or static model, long- or short-run forecasts, theoretical background, inclusion of car use, data requirements, treatment of business cars, car-type segmentation, inclusion of income, of fixed and/or variable car costs, of car quality aspects, of licence holding, of socio-demographic variables and of attitudinal variables, and treatment of scrappage.

Train (1986) and Rand (2002) are both comprehensive surveys of car ownership models, while the literature review here is more focused. More specifically, we cover forecasting models in Great Britain, joint models of ownership and vehicle type/use and different classes of dynamic models. The literature review is organized in two broad categories: static models and dynamic models.

2.1 Static Model

In this section, we first review the "official" car ownership forecasting models in Great Britain, then move on to various advanced discrete choice models. The official forecasting models follow an evolution path of aggregate, partially-disaggregate to disaggregate models[2]. Beyond the simple disaggregate models (typically Multinomial Logit Models), advanced discrete choice models include joint models of car ownership and types (typically Nested Logit Models); models of car ownership and types allowing for heterogeneity (typically Mixed Logit models) and joint estimation of vehicle ownership, types and use (continuous/discrete model).

2.1.1 Aggregate trend extrapolation model (GB: pre 1970s)

In Britain, the very early car ownership forecasts were on the whole unconditional, i.e. they were single-valued estimate without considering the influences of economic and

[2] For a detailed discussion on the history of car ownership forecasting in Great Britain please refer to Whelan (2003).

policy variables (Tanner, 1978). The first formal car ownership forecasting model for Britain is Tanner (1958), which is an aggregate model of trend extrapolation[3].

When applying the extrapolation techniques, it has been recognized that car ownership rates should not increase indefinitely in time due to saturation effects. For this reason, Tanner (1958) pioneered a logistic model that relates car ownership rate (cars per capita C_t) with a time trend t:

$$C_t = \frac{S}{1 + \frac{S - C_0}{C_0} \cdot \exp(\frac{-g_0}{S - C_0} \cdot S \cdot t)} \qquad (1)$$

Where C_0 is the average number of cars per capita in the base year;
g_0 is the marginal growth of average number of cars per capita in the base year (calculated by $\frac{1}{C}\frac{dC}{dt}$ evaluated at t_0);
S is the saturation level.

As a result, knowing C_0 and g_0 in the based year would enable one to extrapolate C_t in future years provided the saturation level S can be estimated (although the estimation of S turned out to be problematic and unreliable).

In response to the criticism of Model (1) being too simplistic, Tanner (1978) extended the trend extrapolation model to include the impact of income and motoring costs:

$$C_t = \frac{S}{1 + \frac{S - C_0}{C_0} \cdot (\frac{I}{I_0})^{-bS} \cdot (\frac{P}{P_0})^{-cS} \cdot \exp[-a \cdot S \cdot (t - t_0)]} \qquad (2)$$

where I is the real GDP per capita, P is the real motoring costs and I_0 and P_0 are the corresponding values in the based year. So, besides the saturation level S, parameters a, b, and c should also be estimated.

[3] Strictly speaking, trend extrapolation models are statistical model rather than econometric model so the distinction between static and dynamic model is moot.

2.1.2 Partially disaggregate model (GB: 1970s and 80s)

In the 1970s, the shortcomings of the aggregate trend extrapolation models were increasingly recognised. The response was to introduce a "partially disaggregate" cross-sectional models while handling the time trend separately using time series approach. Two prime examples are Regional Highway Traffic Model (RHTM, described in Bates, et al. 1978 and cited in Ortuzar and Willumsen, 2001) and 1989 National Road Traffic Forecasts (NRTF).

In RHTM, car ownership was defined as a function of real income deflated by a real car price index, and separate models have been estimated for percentage of households with one plus car ($P1+$) and percentage of households with two plus cars ($P2+$):

$$P_t(1+) = \frac{S(1+)}{1+\exp[-a_1 \cdot (\frac{I_t}{P_t})^{-b_1}]} \quad (3)$$

$$P_t(2+) = \frac{S(2+)}{1+\exp(-a_2 - b_2 \cdot \frac{I_t}{P_t})} \quad (4)$$

The model was calibrated on the basis of national, regional and zonal averages data in Britain for the period of 1979-1975 and hence was a "partially disaggregate" model. The model was found to be stable over time, although the same difficult task of estimating the saturation level remains.

The NRTF (1989) maintained the structure of RHTM. It was supplemented by a "separate identification of time trend" (SIC) model, which use the 1985-1986 National Travel Survey data to establish the effect of income hence isolating the time trend effect.

2.1.3 Fully disaggregate models (GB: 1990s and beyond)

NRTF (1997) was the first fully disaggregate car ownership forecasting model for Great Britain. Car ownership forecasting in the National Transport Model (NTM), currently used by the Department for Transport, have similar structure but include various incremental improvements to NRTF (1997).

- *National Road Traffic Forecasts, 1997*

NRTF (1997) considered five possible methods in car demand forecasting, which include aggregate time series models, joint models of car ownership and use, panel surveys, group cross-section models and individual cross-section models. The method chosen is the individual (household) cross-section model, in which the probability of car ownership at different household income levels is modelled by a logistic function.

In NTRF (1997), two binary models were calibrated for each household type: a P_{1+} model to predict the probability of the household owning at least one car, and a $P_{2+|1+}$ model, defining the conditional probability of the household owning two or more cars, given that they own at least one car:

$$P_{1+} = \frac{S_1}{1+\exp(-LP_1)} \qquad (5)$$

$$P_{2+|1+} = \frac{S_2}{1+\exp(-LP_2)} \qquad (6)$$

The ownership models included saturation levels of maximum car ownership (S_1 and S_2 for 1+ car and 2+|1+ cars respectively), and linear predictors (*LP*) which comprised a linear combination of explanatory variables. The model variables were licences per adult, household income and area type.

- *National Transport Model (Previously known as NRTF 2001)*

Car ownership forecasting in the National Transport Model was described in Whelan (2001, 2003 and 2007) and included various incremental improvements of NRTF (1997). It accounted for the increase in multi-vehicle households by introducing an additional sub-model, which model the conditional probability of a household owning three or more vehicles ($P_{3+|2+|1+}$). Unlike the 1997 NRTF, multiple car ownership by single person households was allowed. To account for the impact of company car ownership on total household car ownership, company car dummies were introduced into the ownership model.

Saturation levels have an important impact upon the results of ownership models. The 1997 NTRF models had allowed variation by household type, but not area type. In the National Transport Model, saturation levels varied according to both household type and area type. Saturation levels were estimated from Family Expenditure Survey data (see Whelan et al. 2000). A general pattern of higher saturation levels in more sparsely populated areas was observed for each model type. Furthermore, a distinct "London" Effect was found, i.e. saturation levels in the Greater London area were lower than in other area types.

2.1.4 Models of Car Ownership and Types

While NRTF (1997) and NTM are fully disaggregate models, they are relatively simple modelling systems, which consist of two or three separate binary choice models (model of Car 1+, Car 2+|1+ and Car 3+|2+|1+). Beyond the simple Binary Logit and Multinomial Logit Models, a number of studies used more advanced discrete choice models to account for correlation between alternatives. In particular, if the household car types are considered in the model, the assumption of independently identically distributed error term in the multinomial logit models is likely to be violated. In this case, the Nested Logit Model, which allows more flexible error structures, becomes a natural candidate. A Nested Logit Model is appropriate when the set of alternatives faced by a decision maker can be partitioned into subsets, called nests, in such a way that: 1) The Independence from Irrelevant Alternative (IIA) property holds within each nest; 2) IIA does not hold in general for alternative in different nests (Train, 2003).

Train (1986) is a modelling system that allows simultaneous estimation of vehicle numbers and types (the model also estimates car use defined as vehicle miles traveled). It has a Nested Logit Structure, where the choice set in the upper nest includes 0, 1 and 2 cars. In the lower nest, the choice set for one vehicle households includes class/vintage of vehicle, while the choice set for two vehicle households includes class/vintage of *pairs* of vehicles. The household chooses the number of vehicles to own and the class/vintage of each vehicle so that the conditional indirect utility function is maximized.

Hensher et al. (1989) is an empirical model of household car holdings in Australia considering both number of cars and types of car owned. It has a Nested Logit structure,

with the vehicle choice decision decomposed into three linked choices: type-mix, body-mix and fleet size (0, 1, 2, 3 or more vehicles). The utility from holding a car bundle can be represented by a conditional indirect utility function, which includes the following variables: consumption prices, qualities, household wealth, expected annualized vehicle capital costs and socio-demographic factors. A particular bundle will be chosen if the utility flows from it exceed the utility obtained from any other car bundles. A car bundle is decomposed into holding of different fleet size, different body types and different models/vintages. The error terms are allowed to be correlated across bundles with different fleet size, body and model/vintage mixed, but are assumed to be IID for bundles with the same choice mixes.

2.1.5 Models of Car Ownership and Types allowing for Heterogeneity

The impacts of household heterogeneity and random taste variations on household car holdings are attracting more and more attention in recent literature. Because of the rapid development in computing technology in recent years, heterogeneous models can now be estimated using simulation without too much difficulty. Among them, Mixed Logit is a highly flexible model that can approximate any random utility model (McFadden and Train, 2000). It allows for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time, which alleviated the three limitation of standard logit. This flexibility gives mixed logit model great advantage in terms of modeling vehicle number and types while allowing for random taste variations. It can also be a highly efficient model in combining revealed preference (RP) and stated preference (SP) data.

In Brownstone et al. (2000), multinomial logit (MNL) and mixed logit models were compared based on data on Californian households' revealed and stated preferences for cars. In the vehicle choice modeling context, they found RP data was critical for obtaining realistic body-type choices and scaling information. SP data was critical for obtaining information about attributes not available in the marketplace, but pure SP models gave implausible forecasts.

Before estimating joint SP/RP models, separate SP and RP models were estimated. The SP models were estimated using both MNL and mixed logit model forms. To identify the presence of significant random error components in the MNL models, the Lagrange

multiplier test from McFadden and Train (1997) was used. Five random coefficients were identified, amongst which four were applied to the different vehicle fuel types modeled, demonstrating large heterogeneity in taste for alternative fuel vehicles. The RP models were also estimated separately. The model result showed that only terms for price and operating cost could be determined with any accuracy due to high co-linearity between vehicle range, speed and acceleration. Joint SP/RP models were then estimated. A scale factor was used to scale the SP data relative to the RP data. While this scale factor is less than one for the MNL model, it is greater than one for the mixed logit model, where the preference heterogeneity is captured by fuel-type error components.

The authors proceeded to make new vehicle forecast for California. The mixed logit models tend to result in higher market shares for the alternative fuel vehicles. A key point here is that the IIA properties of the MNL means a proportionate share of each new vehicle's market share must come from all other vehicles, whereas the mixed logit specification results in the more plausible result that the market share for electric fuel vehicles comes disproportionately from other mini and subcompact vehicles.

2.1.6 Joint Estimation of Car Ownership and Use

While household's car ownership decision is discrete, vehicle use is continuous. A special type of random utility model, joint continuous/discrete model, was developed for this situation (Dubin and McFadden, 1984; Train, 1986). Household's decision on car ownership and use is jointly depicted by a "conditional indirect utility function" and a demand function, whose relationship is established by the so-called "Roy's Identity". The decision-maker chooses the quantities of the continuous goods (e.g. vehicle miles) that maximize his direct utility subject to budget constraint for the given price and income, conditional on choosing a certain discrete alternative (e.g. number of vehicles). A household will only choose for car ownership and drive a positive mileage if the maximum utility of car ownership exceeds the utility of not having a car. Two early examples of joint car ownership and use models are Train (1986) for California and De Jong (1989a, 1989b) for the Netherlands.

Norwegian model system developed by HCG and TØI (1990) is another example. The indirect utility function has two arguments: car use A, measured as the annual driving

distance; and volume of all other goods and services X. The cost of usage is decomposed into fixed costs C and variable cost V. The problem was formulated as:

Maximize {U=U(A,X)}

subject to the budget constraint:

$Y \geq X$, if no car

$Y \geq V_1A_1+C_1+X$, if one car

$Y \geq V_1A_1+C_1+V_2A_2+C_2+X$, if two cars

where Y represents net household income.

If a household does not own a car then it can spend all income on other goods. If the household decides upon car ownership, then to overcome the disutility associated with the fixed costs it must drive a positive number of kilometers. Conditional indirect utility functions were defined for each positive car ownership outcome; for the zero car outcome a direct utility function could be defined. The indirect utility functions give the maximum utility on the budget line and represent the utility of owning a car and driving the optimal distance. The functional form for the demand function for vehicle distance was based upon statistical analysis. The linkage between the indirect utility functions and the demand functions was provided by Roy's identify.

For both cars, significant terms were estimated for the log of remaining household income, the variable cost of driving, the log of household size and percentage urbanization. For the first car only, significant terms were identified for a female head of household; the second car only, significant terms were estimated for age of head of household over 45, and age of head of household over 65.

2.2 Dynamic Models

While the static approach traditionally dominated the study of car ownership and all official forecasting models for Great Britain are static ones, the importance of dynamics is increasingly recognized and many classes of dynamic models have been developed in the last two decades.

2.2.1 Time series Models

Time series model has a distinctive advantage of very light data requirement. Also, it can be quite accurate in terms of short term forecasts. However, these models have a significant drawback: they were unable to include the important influences such as demographic factors. As a result, there are much fewer car ownership models using time series approach. Romilly et al. (1998) and Romilly et al. (2001) are the most notable work in this direction.

Romilly et al. (1998) used a general to specific approach to construct the car demand model, using cointegrating models to establish long term equilibrium conditions (and for long term forecasts) and error correction models to identify short term elasticity. Although the forecasting models appeared to provide plausible income, own-price and cross-price demand elasticities, the actual forecasts were simply unrealistic. Due to the effects of a negative time trend, the forecasted car stock peaked in 2000 and started to decline afterwards.

Romilly et al. (2001) followed their previous research and used five alternative estimation methods to test for cointegrating relationships between per capita car ownership and real per capita personable disposable income, real motoring costs and real bus fares. They are the Engle-Granger two stage, the Phillips-Hansen fully modified, the Wickens-Breusch one-Stage, the auto-regressive distributed lag, and the Johansen maximum likelihood methods. In terms of ex-post forecasting performance the EG2S and ARDL methods gave the best results for car ownership and use respectively, and all four causal models out-perform the ARIMA models. However, in terms of ex-ante forecasts for some 35 years ahead, there was wide divergence in the results between the EG2S/PHFM and WB1S /ARDL methods.

2.2.2 Equilibrium Market Models

Unlike other models that only consider the demand of cars, equilibrium market models are based on the equilibrium mechanism of car demand and supply. There are relatively few equilibrium car ownership models; Cramer and Vos (1985; cited in Rand, 2002) is a good example. The model consisted of two blocks: 1) Car fleet at the end of year t; 2) The market process, with the number of cars purchased and second hand car price determined for each year.

The dynamics of the model laid in the determination of the equilibrium between the number of purchased cars and the supply in the car market. The former was determined by the number of people, the number of households, the average income and the distribution of income, and various prices; the latter was determined by the number of scrapped cars, aging, car bought before. The two unknown endogenous variables in the model were Q_t, the number of purchased cars and $P_{0,t}$, the price of a second hand car (in the supply and demand functions).

The dynamics of the car market was expressed by the adjustments in the demand for the existing number of old cars, via the price of second hand cars, and through its effect on the demand for new cars. So if the price of a second hand car increased, the demand for new cars will increase. The model described the developments from year to year.

The recent study of Meurs et al. (2006) described a dynamic automobile market model for the Netherlands (DYNAMO). The centre of this model was an equilibrium module, where the price mechanism was used to create balance between supply and demand. Unlike the aggregate model of Cramer and Vos (1985), DYNAMO was partially disaggregate, as the model was developed based on 71 types of households. The model also considered 120 separate car types, and the combination of household types and car types described the car ownership in a particular year and thus formed the core of the car ownership model.

2.2.3 Panel Data Model (of car holdings)

Panel data models are disaggregate dynamic models, which can be very efficient as they make the most use of the information embodied in the repeated cross section data. Unlike the time series methods and cross-sectional methods, the panel data method is able to analyse both the cross-sectional and temporal effects. The advantage of panel data over periodic cross section data using different people is that it is possible to control not only for factors which vary across groups of individual, but also for period, age and individual specific effects.

The main restriction for the use of the panel method is the availability of data. In Britain, the only source of national panel data is the British Household Panel Survey

(BHPS), which contains limited transport-related information. Hanly and Dargay (2000) constructed a panel data model using four years (1993-1996) of BHPS data. The authors used the Heckman 2-stage technique to address the attrition problem. The main objective of the study was to examine whether owning a car in the previous year had a significant effect on the current state after controlling for unobserved heterogeneity (true state dependence *vs* spurious state dependence as in Heckman, 1981b). The results strongly supported state dependence but showed that heterogeneity was not significant. The econometric model was a random effect ordered probit model. However, there are two potential shortcomings with the model specification. Firstly, the ordered response choice mechanisms are not consistent with global utility maximization (De Jong et al. 2004), and as identified in Bhat and Pulugurta (1998), the unordered Multinomial Logit Model always has more superior performance. Secondly, the "initial condition problem" is not addressed and the random effect estimator is generally biased due to violation of the orthogonality assumption between the unobserved effects and the explanatory variables. The second point will be explored in greater details in Chapter 7.

Both issues appear to have been addressed in a recent study of Leth-Petersen and Bjorner (2005). It used the panel dataset of 10,565 Danish households between 1992 and 2001, which was created by merging different public administrative registers at individual levels. Their most general model was a mixed logit model, or more specifically, a random effect multinomial logit model that allowed correlation between alternatives. The error term had the structure of $\varepsilon_{it} = C\xi_i + \nu_{it}$, where ν_{it} was a J x 1 vector of residual with IID Gumbel distribution, ξ_i was a J x 1 vector of IID normal parameters and CC' was the J x J covariance matrix of ξ_i (J is the number of alternatives). The initial condition problem was tackled using the approach proposed in Wooldridge (2005), i.e. modelling the distribution of the unobserved heterogeneity conditional on the initial value of the dependent variable y_{i0}. The estimation results showed that the goodness of fit increased as the unobserved heterogeneity and state dependence were introduced in sequence. It also found that the parameters describing the covariance structure became much smaller when state dependence was introduced, suggesting (true) state dependence observed much of the persistence in the data.

However, the model generated very low elasticity, with the income elasticity in the most general model (random effect state dependent model) being merely 0.06.

Nobile et al. (1996) estimated a random effect multinomial probit (MNP) model of car ownership level, using panel data collected in the Netherlands. The data source for the modelling was data drawn from Dutch National Mobility Panel. Ten waves were collected between March 1984 and March 1989. Data from waves 3, 4, 7 and 9, collected between 1985 and 1988, were analysed. In total, the four waves comprised 2,731 households for a total of 6,882 observed choices. The approach used for model estimation was Bayesian: a prior distribution of parameters of longitudinal MNP model was specified and the "posterior" was examined using Markov Chain Monte Carlo methods. It should be noted that such non-parametric Bayesian models are not really suitable for forecasting purpose.

Some other studies used the panel data methods in forecasting based on the international data. Dargay and Gately (1999) based their model on annual data for 26 countries over the period of 1960-1992; Medlock and Soligo (2002) used a panel of 28 countries. Although these studies are particularly useful to identify the similarities and differences among countries, they are aggregate models in essence. Consequently, they have shortcomings similar to aggregate time series models, especially when the emphasis is on individual countries. In an international study of disaggregate household data, Dargay and Hivert (2005) used the European Community Household Panel (ECHP) to investigate car ownership in a number of European countries. The first wave of ECHP included a sample of 60,500 households in 12 EU countries with some countries added or dropped out in later years. However, the effects of dynamics were not investigated in their econometric models.

2.2.4 Pseudo Panel Models

The pseudo panel model is a relatively new approach to car ownership modeling. Dargay and Vythoulkas (1999) was the first study to apply pseudo panel approach to estimate a dynamic model of car based on repeated cross-section survey data. Their econometric models used a pseudo panel dataset constructed from the repeated cross-section data contained in the UK Family Expenditure Survey between 1982 and 1993. The individual households were grouped into cohorts, which were defined in terms of

the year of birth of the household head, and the averages within these cohorts were treated as observations in a panel. By grouping the individual observations into cohorts, one is assuming homogeneity within the cohorts and heterogeneity between the cohorts. Further issues of cohort definition will be discussed in Chapter 3 and 4.

Similar to other empirical studies of pseudo panel models, the measurement error problem was ignored because the number of sample observations in each cohort was sufficiently large. All econometric models had a linear form, among which one fixed effect model (using the Within Estimator) was compared to three "generation models" (OLS, random effect and random effect with first order auto-regressive error). It should be noted that the generation models are in effect restrictive fixed effect models, which constrains the cohort fixed effects to be linear. The estimated coefficients were found to be similar across models and all of them had the expected sign. The generation model was found to have better fit than the fixed effect model.

Being the first study of the kind, Dargay and Vythoulkas (1999) obviously has scope for improvement. For example, the household characteristic variables only included average number of adult and children, which ignored the number of people in work, a variable found to be highly significant in cross sectional studies. While the descriptive data revealed strong "life cycle effects" and "generation effects" of car ownership, only generation effects were considered in the econometric model. Furthermore, all models had a linear functional form, even though the semi-log model is generally found to be a better form for demand function (e.g. semi-log of the income variable implies declining income elasticity).

Some of these issues were subsequently addressed in some follow-on studies. In Dargay (2002), the functional forms investigated included linear, semi-log, double log and log-inverse (another difference was that the cohorts were defined based on both the age of household head and household location). In Dargay (2001), the life cycle effects were captured by five dummy variables representing different age bands. The fixed effect models were found to have better goodness of fit in both studies. They also revealed that the importance of dynamics could be exaggerated in Dargay and Vythoulkas (1999), as the coefficient for the lagged dependent variable was much

bigger in the original study. This finding is supported by the current study, which will be discussed in Chapter 5.

This thesis presents various improvements to the pseudo panel models developed by Dargay and others. It uses a bigger and more recent dataset covering 1982 to 2000. More variables have been used to examine the impacts of household characteristics on car ownership. Two sets of variables have been tested: one including household size plus average number of children and working people per household; the other including the split of eight household types. In estimating the transformed linear models (semi-log and double log), we explore two ways of transformation (average of log or log of cohort average) and their impacts on the modelling results. The life cycle effects are represented by the second polynomial of cohort age rather than dummy variables of cohort age bands. And finally, we have carried out systematic specification search and used the parametric bootstrap techniques to check the robustness of the estimation.

As durable goods, the decision of car ownership for the individual household is clearly discrete. This would raise questions about the appropriateness of linear car ownership models. In any case, it would be beneficial to have a model that is consistent with the microeconomic theory of utility maximisation. For this reason, this study also applies an innovative method that combines pseudo panel with discrete choice model, which enables dynamics and saturation effects to be studies at the same time.

2.2.5 Dynamic Transactions Models

As identified by De Jong and Kitamura (1992), most discrete choice models of household vehicle ownership are vehicle holdings models that describe the likelihood that a household of given attributes will hold a particular set of vehicles. These models quantify the effects on vehicle demand of various vehicle attributes (e.g. price, running costs, make, type, etc.) and household socio-demographic attributes (e.g. income, household size, etc.). On the other hand, dynamic vehicle transactions models view the household vehicles ownership status as a result of a series of transaction decisions to acquire, replace and dispose of household vehicles. They represent changes in a household's vehicle ownership status, such as buying and/or selling of a car. In this way, household car ownership is modeled as a dynamic behaviour process over time.

The transactions choice model for alternative-fuel vehicles in California (Bunch et al, 1995; Brownstone et al 1996) used micro-simulation methods to model dynamics. The household simulation module updated (aged) household by simulating births, deaths, divorces, children leaving home, etc. The transaction timing module took the updated (aged) household and current vehicle holdings as inputs and decided whether or not a vehicle transaction took place during the simulation period, which was set to 6 month to limit the number of transaction to 1. The vehicle transaction was defined to include disposal, replacement and new purchase. If the transaction time module predicted that a vehicle transaction had taken place, the module of transaction type determined exactly what type of transaction took place. The transaction type module used a number of multinomial logit model after the test on the Independence of Irrelevant Alternatives confirmed its suitability. Finally, the household's vehicle holdings were updated after the transactions, and they became the starting values for the next period's simulation.

A more common type of dynamic transactions models is duration models (e.g. Hensher and Mannering, 1994; Gilbert, 1992; De Jong, 1996; Ramjerdi et al. 2000). For example, De Jong (1996) described a disaggregate transactions model system developed and tested by Hague Consulting Group between 1993 and 1995 for the Netherlands. The core of the model system was a duration model which explained the time which elapsed between purchase of a vehicle and its replacement. The Duration decision can be influenced by a number of factors including attributes of the previous car, socio-economic attributes of persons and households, macro-economic development and attributes of the car market. In a duration model, exit from a state is a realization of a stochastic transition process. This process is characterized by a hazard function $h(t)$, which gives the probability of exit from the state immediately after time t, given that the state is still occupied at t. Besides the core duration model, the model system also contained other modules including vehicle type choice models, regression equations for annul use of the present vehicle and module on fuel efficiency.

The simple duration model considers the duration of ownership of vehicle(s) until its replacement (disposal). More recent models consider three types of vehicle transactions: replacing one of the vehicles in the household fleet (replacement), disposing one of the vehicles (disposal) and acquiring a vehicle to add to the fleet (new purchase). The model used is a competing-risks-duration model, where several "latent" hazard

functions describe different ways of exit from the state. The latent hazard that ends the state first will prevail, and other hazards will remain latent. The examples of competing-risks-duration model include De Jong and Pommer (1996), Yamamoto et al (1999) and Mohammadian and Rashidi (2007).

The duration models rely on statistical hazard functions and are not consistent with the micro-economic theory of utility maximization. A small number of studies attempt to bridge this gap and have been developed based on utility maximization theory. A notable example is Golounov et al (2002) and Golounov et al (2004), which used revealed preference data and stated preference data respectively. Their models are based on the intertemporal utility theory (Deaton, 1992), where the decision maker maximizes the intertemporal utility function, which is represented by a discounted sum of utilities in every period. The latter study used the mixed logit model to model random discount rate across individuals, thus accounting for heterogeneity in intertemporal decisions. The major shortcoming of these studies, however, lies in their failure to model the impacts of current choice on future utilities so they can not be regarded as a genuine dynamic model.

Dynamic random utility model explicitly accounts for state dependence. For example, Mohammadian and Miller (2003) used exponentially smoothed weighted average of past choices (Guadagni and Little, 1983) to capture the dependence of current utility evaluations on past transaction choices. This followed the idea that a vehicle transaction itself may have effects on household needs and motivations for automobile ownership level and each transaction can potentially affect the timing and type of the transaction that followed. Heterogeneity across decision makers, on the other hand, was handled by mixed logit formulation. Another truly dynamic theoretical model of car transactions is Adda and Cooper (2000), which used dynamic optimization to investigate the effect of government subsidy on vehicle scrapage. It will be discussed in Chapter 7 with other dynamic models of state dependence so no further details are given here.

2.3 Conclusion

Given the vast number of studies on car ownership, we can not claim the literature review here to be comprehensive. Nevertheless, a few clear patterns still emerge from the review. Firstly, the car ownership models were traditionally dominated by static approach, and it is still the case for the forecasting models in Great Britain. Secondly, dynamic models of car ownership have become a thriving area of research in the past two decades, with many classes of models utilizing a diverge range of theories and methodologies. Thirdly, disaggregate models have become the dominant form of car ownership model, and this is the case for both static and dynamic models.

The trend towards dynamic and disaggregate models puts much heavier requirements on data. Panel data is the preferred form of longitudinal data, but they are difficult and expensive to collect so there are very few high quality panel datasets available. Furthermore, panel data suffer from the problem of attrition, which can be very severe for long running surveys. One way to avoid the collection of expensive panel data is to use a retrospective survey, where the respondents provide information on their vehicle holding and transactions in the past years. This is a common approach used in many dynamic transaction models. However, retrospective survey has a major shortcoming that it can at best collect limited past information of household characteristics and other relevant variables, so most dynamic transaction models have no or very few time-varying covariates (explanatory variables). Another approach to estimate dynamic disaggregate models without the need for panel data is to construct pseudo panels from the rich sources of repeated cross sectional surveys. This is the method adopted in few previous studies and is the main focus of the current project.

Chapter 3 Pseudo Panel Data

The motivation to use the pseudo panel model is to take advantage of the high quality cross sectional survey data available in the UK. The long running Family Expenditure Survey[4] appears to be the best source, which is described in Section One. To construct the pseudo panel, we have to identify which FES variables to be included. In the current study, the main selection criteria are their relevance to the car ownership decision, so we review the factors that influence car ownership levels in Section Two. Section Three discusses the definition of cohort and construction of the two pseudo panel datasets. In Section Four, we examine several pseudo panel variables and they reveal some desired feature of the pseudo panel data. Finally, we describe the aggregate data outside the pseudo panel, which are discussed in Section Five.

3.1 Family Expenditure Survey

In Britain, there are several national surveys containing transport related information. Among them, the longest running and most comprehensive one is the Family Expenditure Survey (FES), which contains a range of variables that are relevant to car ownership modelling. The FES is a voluntary survey of a random sample of private households in the United Kingdom carried out by the Office for National Statistics. It is primarily a survey of household expenditure on goods and services, and household income. The original purpose of the survey was to provide information on spending patterns for the Retail Price Index. Over the years the range of uses has grown and the survey is now multi-purpose. Many previous researches on car ownership modelling in the UK use the FES data as their primary source (e.g. NRTF, 1997; Whelan, 2001; Dargay and Vythoulkas, 1999).

The Family Expenditure Survey is a continuous survey with an annual sample of around 6,500 households. It ran from 1957 to 2001, until it was merged with the National Food Survey to form a new Expenditure and Food Survey. Data is collected throughout the year to cover seasonal variations in expenditures. The FES contains rich

[4] The Family Expenditure Survey is Crown Copy Right material and is obtained with permission from Data Archive.

data on expenditure and income, including vehicle purchasing and servicing costs data. The FES also collects information on socio-economic characteristics of the households, e.g. composition, size, social class, occupation and age of the head of household. Many of these variables have been identified as the important factors influencing car ownership. Table 3.1 shows the data coverage summary of FES.

Table 3-1 Data Coverage Summary of FES

Persons/entities covered:	Households and Individuals
Summary of coverage:	Data Coverage is of household expenditure, income and socio-economic characteristics of households.
Key census variables used:	Age/Date of Birth Ethnic Group Marital status Sex Social Group Socio-Economic Group
Harmonised questions used:	Tenure Type of accommodation Personal characteristics Employment status

(Source: ONS, 2002a)

Regarding the data collection methodology, the fieldwork was carried out by different agencies in Great Britain and Northern Ireland using almost identical questionnaires. Each individual in the household visited aged 16 or over is asked to keep diary records of daily expenditure for two weeks. Information about regular expenditure, such as rent and mortgage payments, is obtained from a household interview along with retrospective information on certain large, infrequent expenditures such as those on vehicles. Regarding the sampling frame, The FES sample for Great Britain is drawn from the Postcode Address File - the Post Office's list of addresses. 672 postal sectors in Great Britain are randomly selected during the year after being arranged in strata defined by Government Office regions (sub-divided into metropolitan and non-metropolitan areas). The Northern Ireland sample is drawn as a random sample of addresses from the Valuation and Lands Agency list.

Besides the Family Expenditure Survey, another possible data source that is suitable for the purpose of the current study is the National Travel Survey (NTS). NTS is a series of household surveys designed to provide regular and up-to-date data on personal

travel and monitor changes in travel behaviour over time. The main advantage of the NTS data is that it includes much more details on vehicle information such as registration details, parking, vehicle subsidies, mileage and fuel. For studies that predict what types of new car might be purchased in the future, these data are essential so the NTS would be the preferable data sources (e.g. Page et al., 2000). However, the main disadvantage of the NTS data is that it has shorter history. It has been running on an ad hoc basis since 1965 and became a continuous survey only since 1988 (ONS, 2001). Since the current study emphasizes a dynamic approach, it is believed that a survey with longer time period would be more appropriate.

In the UK, there are other national databases containing transport related questions, most notably the Census and the General Household Survey. As none of them are adequate for the purpose of the current study, the Family Expenditure Survey has been used as the main data source.

3.2 Factors Influencing Car Ownership

Variables that influence car ownership decisions should be included in the pseudo panel dataset. First of all, it is widely recognized that income is the most significant factor influencing car ownership, and almost all car ownership models in the literature include the income variable in one form or another. The FES data contain information on total household income and disposable household income. Both variables were initially included in the pseudo panel, although only the latter is used in the econometric models since it is generally accepted as a more appropriate measure.

Besides household income, household structure (socio-demographic characteristics) also directly influences car ownership. Average household size, average number of children and people in work per household are all variables found to be relevant to car ownership decisions. The household type influences car ownership level in a similar fashion. In NRTF (1997) and the National Transport Model (NTM), the households are split into eight types based on the number of adults, children and working person in the household. The definition has proved effective in segmentation based on household socio-demographic characteristics and will also be used in the current study. Table 3.2 illustrates the criteria for the eight household types.

Table 3-2 Definition of eight household types

HH type	Description	Defining parameters		
HH 1	One adult, in work	Adult<2	Child=0	Worker>0
HH 2	One adult, not in work	Adult<2	Child=0	Worker=0
HH 3	One adult, with children	Adult<2	Child>0	
HH 4	Two adults, neither in work	Adult=2		Worker=0
HH 5	Two adults, no children	Adult=2	Child=0	Worker>0
HH 6	Two adults, with children	Adult=2	Child>0	Worker>0
HH 7	Three or more adults, no children	Adult>2	Child=0	
HH 8	Three or more adults, with children	Adult>2	Child>0	

Another factor identified by previous studies that influences household car ownership is household location. Accessibility (including the availability and quality of public transport) greatly influences the need for car, but they are very difficult to measure and incorporate in econometric models. Locations are commonly used as proxy for accessibility and have found to be significant explanatory variables in car ownership decisions. The Family Expenditure Survey records the household location as one of five location types, which should be sufficient for our modelling purpose (see Table 3-3 in section 3.4 for the definition of location types).

Finally, car ownership level can be influenced by motoring costs. In some early studies, the total costs of motoring were used as explanatory variables, although in most recent studies the purchase costs and running costs are separated. Some studies also include public transport fares in their econometric models. However, variables of public transports costs are generally found to be insignificant so they are not included in the current study. It should be noted that the available motoring costs data are in the form of aggregate time series.

3.3 Constructing the Pseudo Panel Dataset

The use of pseudo panel data was introduced by Deaton (1985) for the analysis of consumer demand systems. This approach is based on grouping individuals or households into cohorts and thus treating the averages within these cohorts as observations in a panel. In this way, a pseudo panel enables us to follow over time a representative sample of the same cohorts of individuals or households. The pseudo panel approach has been applied not only in microeconomics research, such as study of income and saving (see, for example, Beach and Finnie 2004; Bourguignon et al, 2004;

Baldini and Mazzaferro, 1999), but also in many areas of social science research, including health, education, employment, etc. (e.g. Garner et al., 2002; Glied, 2002; Lauer, 2003; Anderson and Hussey, 2000; Weir, 2003).

To compile a pseudo panel dataset, the cohorts should be defined on the basis of commonly shared characteristics. Such characteristics should be time invariant, such as year of birth of the head of the household, education level, geographic region, etc (Dargay and Vythoulkas, 1999). In the current study, the cohort is defined based on the year of birth of the household head. The choice of the width of the birth cohort is a trade off between the need to have a large number of observations per cohort and the desire to have as much as informative data as possible. The narrower the birth cohort the greater number of birth cohorts and hence the number of data points; on the other hand, this would imply the fewer number of observations per cohort, hence the greater the potential error in estimating the cohort mean (Propper et al. 2001).

The birth cohort is defined in a five-year band in the current study. For example, all the households with its head born between 1901 and 1905 are grouped into a cohort. In 1982, the mean age of household head within this cohort is 79; in 1983, this mean age is 80; in 1984, this mean age is 81, and so on. Likewise, for each sampling year, all the households with its head born between 1906 and 1910 are grouped into a cohort; and for those born between 1911 and 1915, and so on. The objective of such grouping is to track the notionally "same" group of people. Table A.1 in Appendix 1 shows the mean age of all the cohorts constructed in this study. It should be noted that only cohorts with sufficiently large number of observations (more than 100) are included in order to alleviate the measurement error problems.

Furthermore, the FES survey year changed from calendar year to fiscal year since 1994. Since this change would have an impact on the age of the household head, adjustment has been made to allocate each observation into calendar year based on the data collection year. This final wave of the FES data is for year 2000/2001. However, only data for year 2000 are used as there are only a few hundred observations for 2001. In total, the pseudo panel covers 19 years from 1982 to 2000.

The actual construction of the pseudo panel involves importing the various FES data files into an Access database. Using the "Query" tool in Access, the separate data files are first joined up based on common household ID and then different households are aggregated into cohorts based on birth year of household head. Two separate pseudo panel datasets are constructed. The first one uses the entire sample in FES, which is the main dataset and has in total 254 observations from 16 cohorts. It will be used in linear models and nonlinear models predicting the probability of household owning at least one car. The second one uses a sub-sample of car owning households in the FES. It has only 220 observations from 14 cohorts, as more cohort units have to be discarded due to small sample size. It will be used in nonlinear models predicting the probability of household owning two or more cars conditional on ownership of the first car.

3.4 Examining the Pseudo Panel Variables

The constructed pseudo panel data set contains 28 primary variables, which were directly derived from the FES. They fall into five categories, which are summarized by Table 3.3. The variables refer to the average characteristics of all household samples within each cohort. It should be noted that car ownership in the current project is defined based on FES variable A160 "Cars Owned or Used"[5], which includes both privately owned and company cars but excludes light goods vehicles such as vans.

Some of the variables show strong trends across age of cohorts and time[6]. For example, the household size increases as the age of the household head increases up to around 40, and then starts its steady decline. The average household income reaches its peak when its head is in his late 40s. In the following section, three selected variables will be discussed in further detail to reveal the particular cohort age and time trend effects. On the other hand, residence area data are more or less random across cohorts. Regarding percentage of households living in Greater London (Area 1), it varies between 3% and 20%; regarding percentage of households living in the least populated rural area (Area 5), it varies between 10% and 34%.

---

[5] Other FES variables (not used) include A149 "Cars owned in household" and A143 "Number of cars & vans currently owned".

[6] All discussions in this section are based on the primary dataset based on full FES sample, although the results are similar for the second one of car owning households.

Table 3-3 Variables in the Pseudo Panel Dataset

Category	Variable	Description
Transport data	Car	Number of cars owned or used by household
	R_{1+}	Percentage of household owning at least one car
	R_{2+}	Percentage of household owning two or more cars
	PTExp	Average weekly public transport expenditure per person
Income and expenditure data	Inc_{toal}	Weekly household income
	Inc	Weekly household disposable income
	Exp	Weekly household expenditure
Household demographic data	HHSize	Household size
	Child	Average Number of Children
	Adult	Average Number of adult
	Worker	Average Number of working persons
	HH1	Percentage of household as type 1[7]
	HH2	Percentage of household as type 2
	HH3	Percentage of household as type 3
	HH4	Percentage of household as type 4
	HH5	Percentage of household as type 5
	HH6	Percentage of household as type 6
	HH7	Percentage of household as type 7
	HH8	Percentage of household as type 8
Residence area data	Area1	Percentage of household living in Greater London Area
	Area2	Percentage of HH living in metropolitan districts and central Clydeside conurbation
	Area3	Percentage of HH living in areas with a population density of 7.9 or more persons per hectare
	Area4	Percentage of HH living in areas with a population density of 2.2 to 7.9 persons per hectare
	Area5	Percentage of HH living in areas with a population density of less than 2.2 persons per hectare
General data	Year	Year
	Cohort	Cohort ID (1 to 16)
	Age	Average age of household head in a cohort
	Count	Number of observations within cohort

3.4.1 Number of Cars Owned or Used by the Household

The pseudo panel data clearly show the difference of car ownership between cohorts (which have different ages) and between years. Figure 3.1 compares the number of cars owned or used by households for two cross sections of cohorts in 1982 and 2000. In a given year, the average car ownership is low for both the old and young cohorts and the car ownership is the highest for the mid-aged cohort (household head in late 40s). In 1982, the cohort with the highest car ownership was the one with head of household

[7] For the definition of household type see Table 3.2.

born between 1931 and 1935, i.e. aged between 47 and 51. In 2000, the cohort with the highest car ownership was the one with household head born between 1951 and 1955, i.e. those in a similar age band between 45 and 49. However, the maximum (average) car ownership was 1.38 in 2000, significantly higher than that of 1.11 cars in 1982.

Figure 3-1 Average Number of Cars for two Cross Sections of Cohorts: 1982 and 2000

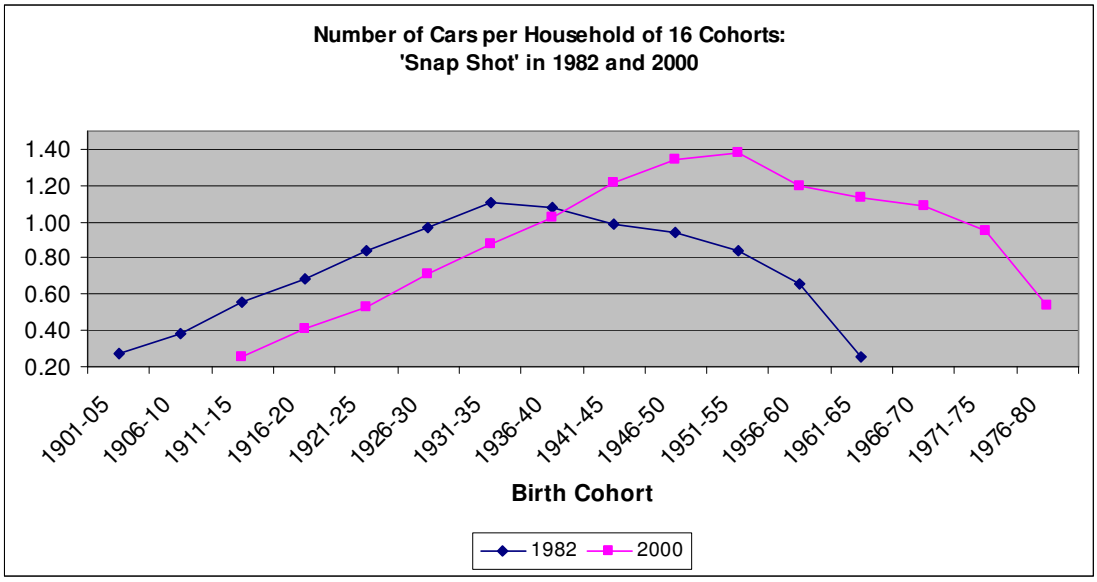
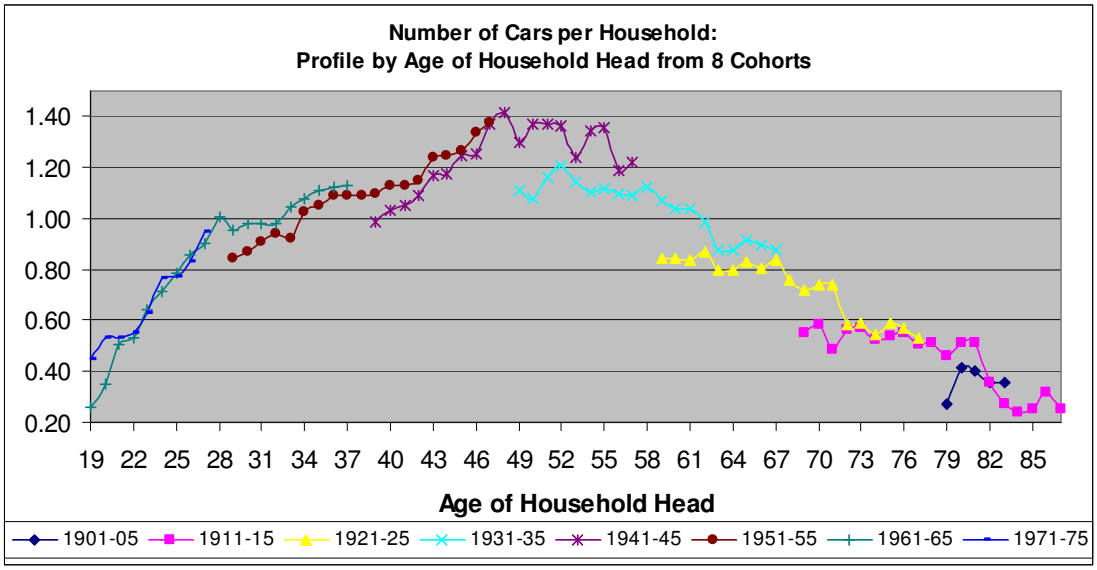


Figure 3.2 presents the car ownership of eight selected cohorts. As cohorts age over time, the mean age of the household head in these eight cohorts covers an entire range between 19 and 85 over the sample period (with overlaps between cohorts). By plotting the number of cars owned against the mean age of the household head, we are able to see the change of car ownership over time for each cohort and make comparison between different cohorts.

There are two apparent trends. First, by combining all the eight cohorts, the trend shows that car ownership rises and falls according to the age of household head, with the peak of 1.42 cars per household when the head is 48 years old. Second, by comparing the car ownership figures of the adjacent cohorts, the trend shows that for any given age, the cohort with younger household head tends to have higher car ownership. These two trends were referred as “life cycle effect” and “generation effect” in Dargay and Vythoulkas (1999). They also found that the difference amongst generations appeared to be declining for the most recent generations. Using more

recent data, the current study found that this "diminishing generation effect" is more apparent.

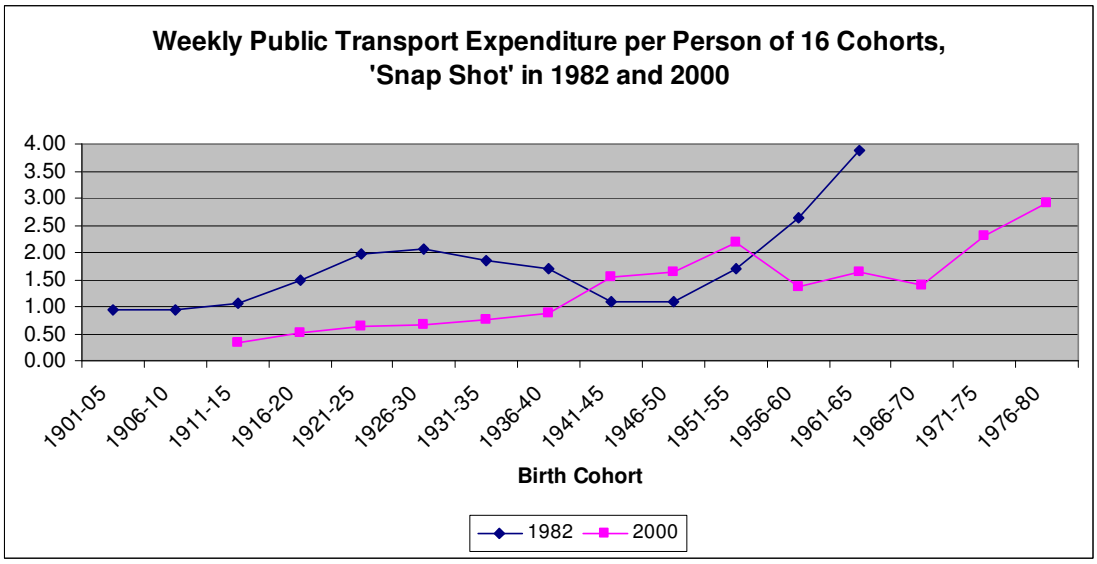
Figure 3-2 Average Number of Cars per Household, Profile by Age of Household Head from Eight Cohorts



3.4.2 Average Weekly Public Transport Expenditure per Person

The average public transport (PT) expenditure per person within the household also varies according to the age of household head. Figure 3.3 shows the different PT expenditure per person for two cross sections of cohorts in 1982 and 2000.

Figure 3-3 Average Weekly Public Transport Expenditure per Person, 1982 and 2000

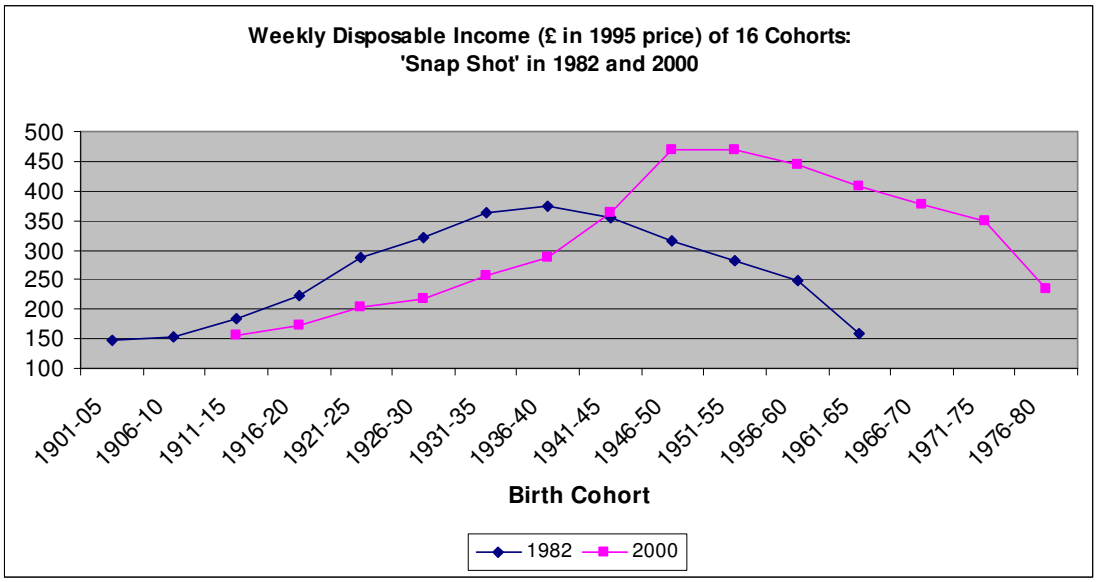


Both years' data reveal a similar trend. It is clear that the PT expenditure per person is the highest in the youngest cohort. This is not surprising given the low car ownership within this cohort. There is a local maximum of PT expenditure per person for mid-aged cohorts (in early 50s). This could be because households within this cohort are likely to have grown-up children, who are not yet car owners and have to pay higher (or full) fares for public transport. The average PT expenditure per person is low for old cohorts, since most old aged pensioners either travel free or pay concessionary fares on many public transport services.

3.4.3 Weekly Household Disposable Income

The weekly household disposable income also shows a strong trend across cohorts. Firstly, we present the profile of household income for two cross sections of cohorts in 1982 and 2000 (Figure 3.4). It shows that the old and young cohorts have lower disposable income, while the mid-aged cohorts have the highest income level. This trend is very similar to that of car ownership, suggesting high correlation between car ownership and income. Since all the expenditure and income data in the pseudo panel dataset have been converted to 1995 prices based on Retail Price Index, Figure 3.4 also shows the real increase of income level for similar age group from 1982 to 2000.

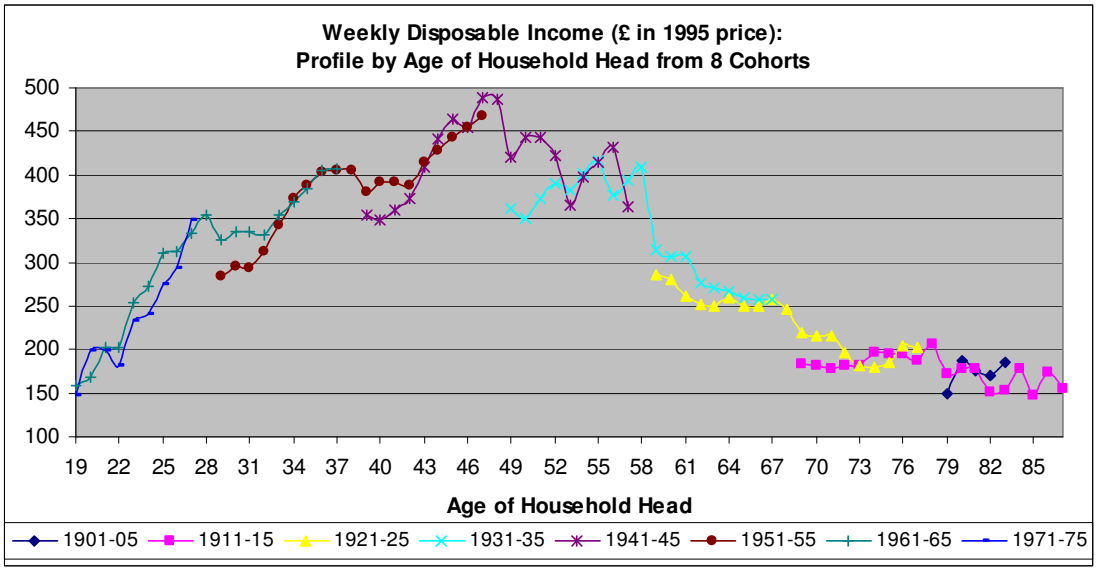
Figure 3-4 Household Weekly Disposable Income, 1982 and 2000



It is revealing to track the change of household income level according to the age of household head for the eight selected cohorts. Figure 3.5 shows that weekly household

disposable income rises as the age of household head increases and reaches its peak when the household head is in late 40s. For the cohort whose household head is born between 1941 and 1945, the weekly disposable income is the highest of £490 when its household head is aged 47.

Figure 3-5 Weekly Disposable Income: Age Profile from Eight Cohorts



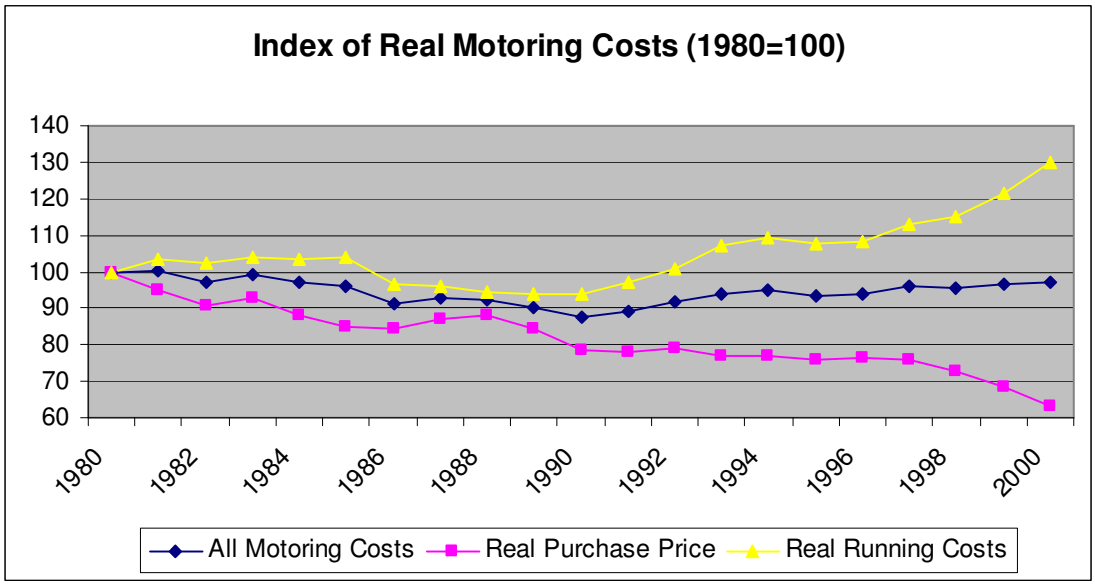
3.5 Aggregate Time Series Data

Certain time series data from various sources are also used in this study. They include the motoring costs and demographic data. The motoring costs directly influence car ownership so they have to be included in the econometric model. Past aggregate demographic data are not directly used in model estimation, although they help to establish future trends, which will be useful for forecasting.

3.5.1 Motoring Costs

The motoring costs data include the index of real car purchase price and real car running costs, which were supplied by the Department for Transport and re-produced in Whelan (2003). For comparison, we also obtain the index of all (real) motoring costs from the Department for Transport publication "Transport Trends" (DfT, 2004). Figure 3-6 shows the three indices of real car purchase price, running costs and total costs between 1980 and 2000, which covers the entire modelling period in the current study.

Figure 3-6 Index of Real Motoring Costs: 1980-2000 (1980=100)

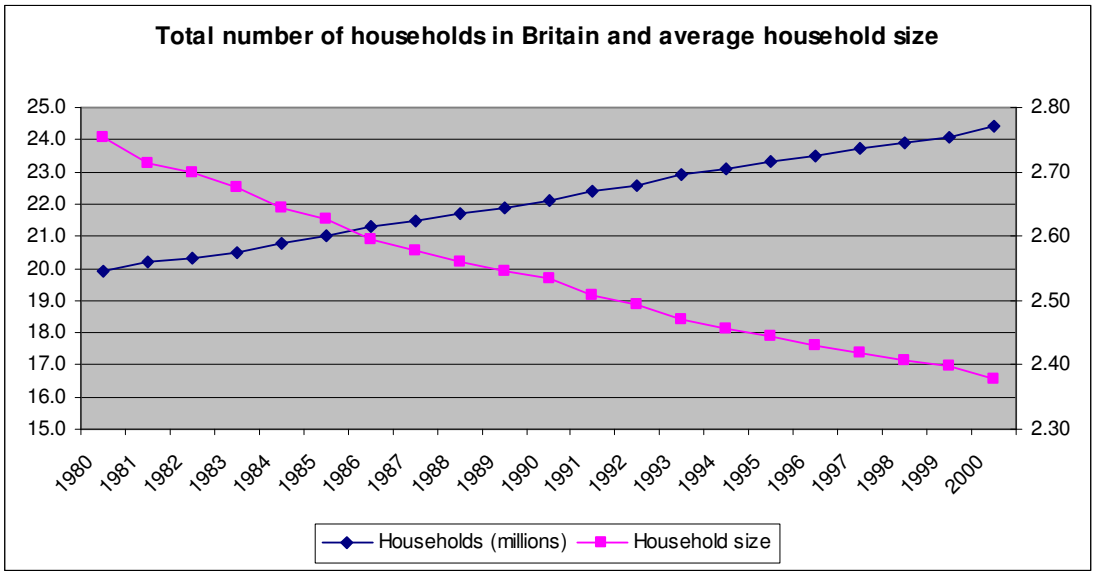


There is significant reduction of real purchase price over time and the index is reduced by about 37% in 20 years. On the other hand, there is only small fluctuation of real running costs index before 1995, and then the costs jump up by over 20% in 5 years. This rapid rise of real running costs happens at the same time as the significant increase of multiple car ownership, which results in a running cost coefficient with unexpected sign in some of the econometric models (detailed results will be presented in Chapter 7 and 8). Overall, the total (real) motoring costs barely change over the 20 year period.

3.5.2 Demographic Data

Figure 3-7 presents the total number of households in Great Britain and average household size between 1980 and 2000. Over the period, the total number of households is increased by 23%, while the average household size is reduced by 14%. The population change over the same period is a more modest 6% increase. While these data are not used in the estimation of the econometric models, it is important that the trend in the change of demography is reflected at the forecasting stage.

Figure 3-7 Total Number of Households and Average Household Size: GB 1980-2000



(Sources: Derived from ONS, 2002b: "Population Trends", 107, Spring 2002)

3.6 Conclusion

This Chapter describes the construction of the pseudo panel. The cross sectional data used are the Family Expenditure Surveys between 1982 and 2000. The pseudo panel includes variables that influence car ownership and hence are required for the econometric modelling. The cohorts are defined based on the year of birth of the household head, and two pseudo panel datasets have been constructed, one from the full FES sample while the other from a sub-sample of car owning households. Further examination of car ownership and income variables reveals a clear "hump" shape life cycle, which is similar to other pseudo panel studies. As will be shown in the next chapter, to minimize the measurement error problem, the cohort should be defined in a way such that the population cohort means of the variables concerned vary as much as possible over time. The life cycle profiles of these variables show that these conditions are likely to have been met.

Chapter 4 Measurement Error and Linear Static Fixed Effect Model

After the construction of the pseudo panel data set, the next step is to estimate models with a linear functional form. We start by investigating the consistency of pseudo panel estimators. Although the primary interest of using the pseudo panel approach is to explore the dynamic effect, the issue of consistency can first be explored within the context of static model. The first section of this chapter explores the link between the estimators based on micro survey data and pseudo panel data. It shows that the Weighted Least Square Estimator based on cohort means is equivalent to the Instrumental Variable (IV) estimator based on individual data in the micro survey and using cohort dummy as instruments. The second part of this chapter discusses three pseudo panel estimators, which consider the observation in the pseudo panel dataset as an error-ridden cohort average. The third section presents the conditions required to ignore the measurement error problem. The fourth section reports the empirical results of linear static car ownership model, estimated using the constructed pseudo panel. Two sets of results are reported, one based on Weighted Least Square Estimator, while the other one treats the dataset as a real panel and investigates various panel estimators including fixed effect, random effect and heterogeneous models. Each includes systematic specification search and discussion of estimation results. The last section is a brief conclusion.

4.1 Weighted Least Square Estimator

The panel data models have the advantages of being able to consider fixed individual effects and dynamic effects. However, the limited availability of the panel data sets, together with the attrition problem for the existing few, constrains the practical application of such models. One alternative approach, as suggested by Browning et al (1985) and Deaton (1985) and adopted in this study, is to estimate fixed effect model based on cohorts rather than individuals. Using a continuous survey that generates random sample of the population in every year, the cohort means can be calculated from each sample and followed through time.

In this chapter, we will show that the Weighted OLS estimator using cohort means is equivalent to an Instrumental Variable estimator using individual data. It is easier to start from considering the individual economic relationship. Assuming such relationship is linear (in the parameters, though not necessarily in the data)

$$y_{i(t)t} = x'_{i(t)t}\beta + \varepsilon_{i(t)t} \qquad (1)$$

for individual i sampled in year t, noting that the individuals are different over the cross-sections; $x_{i(t)t}$ is a K by 1 vector that may contain the dummy variables indicating each cohort.

Assuming that there exists a vector of time-invariant instrumental variables, $z_{i(t)}$, satisfying the standard Instrumental Variable condition including:

$$E\{\varepsilon_{i(t)t} z_{i(t)}\} = 0 \qquad (2)$$

Writing (1) in matrix form and projecting the columns of X in the column space of Z, we have an exact-identified set of Instrumental Variables $\tilde{X}$:

$$\tilde{X} = Z(Z'Z)^{-1}Z'X \qquad (3)$$

This leads to the standard Instrumental Variables Estimator b_{IV}:

$$b_{IV} = (\tilde{X}'X)^{-1}\tilde{X}'y \qquad (4)$$

By noting $M_z = Z(Z'Z)^{-1}Z'$, (4) can be written as:

$$b_{IV} = (X'M_z X)^{-1}(X'M_z y) \qquad (5)$$

By selecting appropriate instrumental variables $z_{i(t)}$, the link between the estimator based on micro survey data and that based on aggregate cohort data can be established. If the instrumental variables are the dummy variables indicating the mutually exclusive cohorts, i.e. $z_{i(t)=}z_c$, it follows:

$$M_z = \begin{bmatrix} M_{1,1} & \dots & 0 & 0 & \dots & 0 \\ & & \dots & & & \\ 0 & \dots & M_{c,1} & 0 & \dots & 0 \\ 0 & \dots & 0 & M_{1,2} & \dots & 0 \\ & & \dots & & & \\ 0 & \dots & 0 & 0 & \dots & M_{c,t} \end{bmatrix} \qquad (6)$$

Here, $M_{c,t}$ is an idempotent matrix for cohort c in cross section t,

$$M_{c,t} = \frac{1}{n_{c,t}} i_{c,t} \cdot i'_{c,t} \qquad (7)$$

where $n_{c,t}$ is the number of sample observations and $i_{c,t}$ is the $n_{c,t}$ dimensional vector of ones.

Similarly, write X and y in (5) (pooled cross sections) as:

$$X = \begin{bmatrix} X_{1,1} \\ \cdots \\ X_{c,1} \\ \cdots \\ X_{c,t} \end{bmatrix}; \; y = \begin{bmatrix} y_{1,1} \\ \cdots \\ y_{c,1} \\ \cdots \\ y_{c,t} \end{bmatrix}$$

And substituting (6) and (7) into (5), it is straightforward to show that

$$b_{IV} = (\sum_{TC} n_{c,t} \bar{x}_{c,t} \bar{x}'_{c,t})^{-1} (\sum_{TC} n_{c,t} \bar{x}_{c,t} \bar{y}_{c,t}) \quad (8)$$

where $\bar{x}_{c,t}$ is a K by 1 vector representing the average of $x_{i(t)t}$ for cohort c in year t; while $\bar{y}_{c,t}$ is a scalar representing the average of $y_{i(t)t}$ for the corresponding cohort. TC is the total number of cohorts over all the sample years, i.e. $TC = C \cdot T$.

Note that the Ordinary Least Square (OLS) estimator based on the cohort average can be expressed in a standard form as (9) (assuming a pseudo panel dataset has been compiled, this means directly applying OLS to the pseudo panel:

$$b_{OLS} = (\bar{X}'\bar{X})^{-1} \bar{X}'\bar{y} \qquad (9)$$

If each of the observations in (9) is weighted by the square root of $n_{c,t}$, it can be re-written in the form of (8). This result show that the Weighted OLS estimator based on cohort average is equivalent to the IV estimator based on micro survey data.

The above analysis has two important implications for the pseudo panel estimation, if we assume the economic relationship between the dependent variable and explanatory variables is linear and holds for individuals. The first is that any linear transformation needs to be done on the micro survey data, and the variables in cohort model would be the average of the transformed data, e.g. using average of log income rather than log of average income. The second is that each observation in the pseudo panel needs to be weighted by the square root of the sample size of the cohort.

4.2 Consistent Estimation of FEM with Measurement Error

In a pseudo panel, each of the observations represents the sample average of a particular cohort in one year. The deviation of sample means from the true cohort means in the population is the measurement error, which would result in biased OLS estimation. Such a problem is likely to be acute if the sample size is too small or the mean is skewed by any extreme numbers. Since the micro data from surveys are used to construct the sample cohort means, they can also be used to construct estimates of the variances and covariances of the sample means (Deaton, 1985). Consequently, a number of Error-in-Variable Estimators have been proposed in the literature to consistently estimate the population relationships.

Assuming the economic relationship for the observed sample means for cohort *c* in year *t* is:

$$\bar{y}_{ct} = \bar{x}'_{ct}\beta + \bar{\lambda}_{ct} + \bar{\varepsilon}_{ct} \qquad c = 1, \ldots, C; t = 1, \ldots, T \qquad (10)$$

As $\bar{\lambda}_{ct}$ is the average of fixed effect for members of cohort *c* that are sampled in year *t*, it is not constant over time. Moreover, $\bar{\lambda}_{ct}$ is unobservable and generally correlated with $\bar{x}_{ct}$. As a result, the within estimator based on (10) will generally be biased. In this case, we need to consider the economic relationship in the cohort population:

$$y^*_{ct} = x^*_{ct}{}'\beta + \lambda_c + \varepsilon^*_{ct} \qquad c = 1, \ldots, C; t = 1, \ldots, T \qquad (11)$$

where y^*_{ct} and x^*_{ct} are cohort population means, which are unobservable; λ_c is the cohort population fixed effect, and if we assume close cohort (no birth or death of cohort members), it will be constant over time. Note that (11) is an aggregated version of equation (1), with $x_{i(t)t}$ in (1) containing cohort identifying dummy variables. Comparing (10) and (11), it is clear that $\bar{x}_{ct}$ and $\bar{y}_{ct}$ are error-ridden estimate of y^*_{ct} and x^*_{ct} (note that $\bar{x}_{ct}$ and $\bar{y}_{ct}$ have to be weighted by the square root of sample size in each cohort). It is common to assume the measurement errors follow independent identical distribution in the literature (e.g. Deaton, 1985; Verbeek and Nijman, 1993; Biorn, 1992; Marshall, 1992; Devereux, 2003):

$$\begin{pmatrix} \bar{y}_{ct} - y^*_{ct} \\ \bar{x}_{ct} - x^*_{ct} \end{pmatrix} \sim iid\left(0, \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix}\right) \qquad (13)$$

where σ_{00}, σ and Σ can all be consistently estimated from the individual observations and are assumed to be known here.

Write the pseudo panel within estimator of (10) as:

$$\beta_{within} = M_{xx}^{-1} m_{xy} \qquad (14)$$

where: $$M_{xx} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \qquad (15)$$

$$m_{xy} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c)' \qquad (16)$$

with $\bar{x}_c$ and $\bar{y}_c$ being the mean of cohort c over T years. Also, note the population counterpart of (15) as:

$$M_{xx}^{*} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(x_{ct}^{*} - x_c^{*})(x_{ct}^{*} - x_c^{*})' \qquad (17)$$

with probability limit of: $$\operatorname*{plim}_{CT\to\infty} M_{xx}^{*} = \Omega \qquad (18)$$

Based on assumption (13), Verbeek and Nijman (1993) showed that:

$$\operatorname*{plim}_{CT\to\infty} M_{xx} = \Omega + \frac{T-1}{T}\Sigma \qquad (19)$$

and similarly:

$$\operatorname*{plim}_{CT\to\infty} m_{xy} = \Omega\beta + \frac{T-1}{T}\sigma \qquad (20)$$

From (14), (19) and (20), it is clear that the pseudo panel within estimator β_{within} is biased:

$$\operatorname*{plim}_{CT\to\infty} \beta_{within} = (\Omega + \frac{T-1}{T}\Sigma)^{-1}(\Omega\beta + \frac{T-1}{T}\sigma) \qquad (21)$$

This leads to two simple Error-in-Variable Estimators proposed in the literature. Deaton (1985) assumed $T \to \infty$, and the resulting unbiased estimator (noted as *EVE1*) can be written as:

$$\tilde{\beta}_{EVE1} = (M_{xx} - \Sigma)^{-1}(m_{xy} - \sigma) \qquad (22)$$

While Verbeek and Nijman (1993) assumed $C \to \infty$, leading to unbiased estimator (*EVE2*) of:

$$\tilde{\beta}_{EVE2} = (M_{xx} - \frac{T-1}{T}\Sigma)^{-1}(m_{xy} - \frac{T-1}{T}\sigma) \qquad (23)$$

As suggested by Verbeek and Nijman (1993), (22) and (23) can be generalized to a class of Error-in-Variable Estimator (EVE_a), with different weight a attached to the standard deviation of cohort measurement error Σ and σ:

$$\tilde{\beta}_{EVE_a} = (M_{xx} - a\Sigma)^{-1}(m_{xy} - a\sigma) \qquad (24)$$

In this case, the Within estimator, *EVE1* and *EVE2* can all be seen as special case of EVE_a, with a being 0, 1, and (T-1)/T respectively.

Also in Verbeek and Nijman (1993), the authors proposed a Mean-Square-Error estimator that is optimal in finite samples. They argued that it may be advantageous to choose a that minimized MSE, even though the implied estimator will suffer from inconsistency. More specifically, if the fixed individual effects (λ_c) are uncorrelated with the explanatory variables ($\bar{x}_{ct}$), minimum MSE is obtained for a = 0, the within estimator on the pseudo panel of cohort data. If there is correlation between the individual effects and the explanatory variables, the optimal a is a value between 0 and (T-1)/T, depending on many factors such as number of sample years, cohort sample sizes, correlation between λ_i and x_i, extent of variation of cohort means over time, etc.

In the more recent work by Devereux (2003), the author established the exact equivalence between Error-in-Variable Estimator and Jackknife Instrumental Variables Estimator (JIVE). Based on earlier work in the JIVE literature, Devereux showed that the approximate bias of JIVE, and hence EVE, is proportional to (-K-1), with K as the number of explanatory variables. Consequently, he introduced an "Unbiased EVE Estimator", with weight a = (CT – K – 1)/(CT). Such estimator is approximately unbiased to order 1/(CT), which may be particularly useful in many practical applications when there are only a small number of cohorts and it is not appropriate to assume $C \to \infty$.

4.3 Conditions to Ignore Measurement Error Problem

Although various Error-in-Variable Estimators have been proposed, the majority of empirical work uses standard within estimator. The main defense of such practice is

that the measurement error problem can be largely ignored if the number of sample observation in a cohort is sufficiently large. It has been shown that such defense is more or less justified if the number of observation in a cohort is greater than, say, 100, although it is equally important that the true cohort means vary over cohorts and/or over time (Verbeek and Nijman, 1992).

Assuming the data generating process for each individual in the *population* is as follows:

$$y_{it} = x_{it}\beta + \lambda_i + \varepsilon_{it}, \quad t = 1, \ldots, T \qquad (25)$$

where x_{it} is a scalar in this simplified case. λ_i is the individual effect (fixed over time), and it may be correlated with x_{it}. Verbeek and Nijman (1992) have shown that the asymptotic bias of the pseudo panel within estimator is related to how the cohorts are defined. In particular, the authors assumed that cohorts are defined on the basis of a continuous variable z, which satisfies:

a) z is distributed independently across individuals with variance normalized to one;
b) the support of the density of z is split into C intervals with equal probability mass, each intervals corresponding to a particular cohort;
c) the correlation between explanatory variable x_{it} and z_i is as:

   $$x_{it} = \mu_{it} + \gamma_t z_i + \upsilon_{it} \qquad (26)$$

   where $E\{\upsilon_{it} \mid z_i\} = 0$ and (for mathematical convenience) equicorrelation of υ_{it} over time, i.e. $\mathrm{cov}\{\upsilon_{it}, \upsilon_{is}\} = \rho\sigma_\upsilon^2$.

Following Mundlank (1978) and Chamberlain (1984), Verbeek and Njiman further assumed the individual effects λ_i are correlated with the x's in the following way:

$$\lambda_i = \kappa\bar{x}_i + \xi_i \qquad (27)$$

where $E\{\xi_i \mid x_{it}\} = 0$ for all $t = 1, \ldots, T$ and $\bar{x}_i = (1/T)(x_{i1} + \ldots + x_{iT})$.

After some fairly involved mathematical manipulation, the authors have shown that under the above assumptions the asymptotic bias of the pseudo panel within estimator $\tilde{\beta}_{within}$ is:

$$\underset{C\to\infty}{\text{plim}}(\tilde{\beta}_{within}-\beta)=\kappa\cdot[\frac{1+(T-1)\rho}{T}]\cdot\frac{\tau\omega_2}{\omega_1+\tau\omega_2}=\delta_{\max}\cdot\frac{\tau\omega_2}{\omega_1+\tau\omega_2} \tag{28}$$

where τ = (T - 1) / T; ω_2 is the measurement error variance in $\bar{x}_{ct}$; ω_1 is the true within cohort variance in the population. More specifically:

$$\omega_2=\underset{C\to\infty}{\text{plim}}\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct}-x_{ct}^{*})^2=\frac{\sigma_v^2}{n_c} \tag{29}$$

with n_c the number of individuals in cohort c; and

$$\begin{aligned}\omega_1&=\underset{C\to\infty}{\text{plim}}\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(x_{ct}^{*}-\bar{x}_{c}^{*})^2\\&=\frac{1}{T}\sum_{t=1}^{T}(\mu_t-\frac{1}{T}\sum_{s=1}^{T}\mu_s)^2+\frac{1}{T}\sum_{t=1}^{T}(\gamma_t-\frac{1}{T}\sum_{s=1}^{T}\gamma_s)^2\end{aligned} \tag{30}$$

with $\bar{x}_{c}^{*}=\frac{1}{T}\sum_{t=1}^{T}x_{ct}^{*}$.

When neither μ_t nor γ_t varies with t, the true within cohort variances ω_1 is zero, so the bias in the within estimator is maximal; in another word, the increase of cohort size will not reduce the bias, which is always δ_{max}. When $\omega_1 > 0$, the choice of larger cohorts will lead to smaller ω_2, thus reduce the asymptotic bias. For example, if $\omega_1/\sigma_v^2=0.5$, increase the cohort size from 50 to 100 will reduce the bias from 3.8% to 2% of δ_{max}; if $\omega_1/\sigma_v^2=0.05$, similar increase of cohort size will reduce the bias from 29% to 17% of δ_{max}. On the other hand, the cohort sizes may be chosen smaller if the cohort identifying variable is chosen in such a way that true within cohorts variance (ω_1) is large relative to σ_v^2.

The above discussion shows that how the cohorts are constructed has direct implication on the bias of the within estimator if the pseudo panel is to be treated as genuine panel. Cohort identifying variables should be chosen in a way that maximizes true within cohort variance; in another word, the population cohort means should vary as much as possible over time. As the true population cohort means are not directly observable, the variation of sample cohort means might give some useful indications. The discussion in Chapter 3 shows clear "life cycle effect" and "generation effect" in the explanatory variables, which indicates variation of cohort means across time and cohorts.

Nevertheless, it also requires the sample number in each cohort to be sufficiently large, otherwise the observed variation of sample cohort means could be purely due to sampling errors. During the construction of the pseudo panel, only cohorts with 100 or more observations are included, thus alleviating the problem of sampling errors. It should be stressed that the bigger the cohort sample size, n_{ct}, the smaller the bias of the pseudo panel within estimator. All these discussions suggest that it might be appropriate to ignore the measurement error problem and avoid using Error-in-Variable estimators.

4.4 Empirical Results from Static Car Ownership Model

The estimation of car ownership model is based on the pseudo panel dataset constructed from the Family Expenditure data, which contains 254 observations, covering 16 cohorts for the period of 1982-2000. The dependent variable is average number of cars per household in cohort c in year t ($A_{c,t}$). The explanatory variables are those that influence household car ownership level, as identified in the literature: income ($I_{c,t}$), household structure or demographic characteristics ($S_{c,t}$), age of household head ($G_{c,t}$), household locations ($L_{c,t}$) and motoring costs (M_t). Equation 31 represents the empirical models to be investigated:

$$A_{c,t} = f(I_{c,t}, S_{c,t}, G_{c,t}, L_{c,t}, M_t) + \varepsilon_{c,t} \qquad (31)$$

The following are detailed descriptions of the explanatory variables. The income variable is noted as "Inc", which refers to the average household disposable income within each cohort, deflated by RPI.

Regarding the household structure variables, the "Child" variable refers to the average number of children per household within each cohort, with others defined similarly. As described in the previous chapter, there is an eight way categorization of the household type based on the number of children, adults and working persons in the household. The variables showing proportion of household within each category will also be included in some of the models as alternative representation of household demography. To control for the "generation effect" and household characteristics not captured by the household structure variables, second polynomial of the average age of household head in each cohorts (variable "Age" and "AgSq") are included in the regression.

The "Area" variables refer to percentage of household living in Greater London, Metropolitan districts/Central Clydeside Conurbation and other areas with various population densities. In some models, Area1 and Area2 are combined to form a new variables, "Met", which refers to the proportion of household living in all metropolitan areas. To correspond with the "Met" variable, Area5 is also called "Rural" in some models, referring to the least populated rural areas[8].

Regarding motoring costs, the index of car purchasing costs and car running costs have been deflated by RPI to obtain the real purchasing costs (variable "Price") and real running costs (variable "RunCst"). Note that these two variables do not vary between cohorts for any specific year.

Finally, the prefix of "Ln" represents the logarithmic transformation of a variable. For example, variable "LnInc" would represent the log of average disposable income of a cohort. As discussed in the previous section, the Weighted Least Square estimator of the pseudo panel requires the linear transformation to be carried out at the individual data before averaging. To differentiate this, the prefix of "ALn" is used to identify the average of logarithmic instead of logarithmic of cohort average. For example, variable "ALnInc" would represent the average of the log transformed household disposable income in a cohort.

Table 4.1 summarizes the key descriptive statistics of the dependent and selected independent variables. The average number of car owned or used by household is 0.86, while the maximum and minimum is 1.42 and 0.19 respectively. The average size of household is 2.40, and the average number of employed people in the household is 1.05. The real household disposable income (in 1995 prices) varies between £147 and £490 per week.

A systematic specification search has been carried out to determine the model with best fit. As discussed in the previous section, the measurement error problem can be ignored

[8] For detailed description of the household structure and location variables refer to the previous chapter.

Table 4-1 Descriptive statistics of the variables

	Car	Inc	Child	Adult	Worker	HHSize	Area1	Area2	Area3	Area4	Area5
Median	0.91	302.49	0.37	1.81	1.27	2.27	0.10	0.22	0.22	0.22	0.24
Mean	0.86	298.04	0.55	1.84	1.05	2.40	0.10	0.22	0.22	0.21	0.24
Stdev	0.31	96.52	0.57	0.26	0.65	0.69	0.03	0.03	0.03	0.03	0.04
Max	1.42	489.56	1.85	2.50	2.16	3.87	0.20	0.32	0.37	0.28	0.34
Min	0.19	147.09	0.00	1.22	0.03	1.22	0.03	0.15	0.16	0.11	0.10

if the number of observations in each cohort is sufficiently large ($n_{ct} \to \infty$). The current dataset is deemed to satisfy this requirement so no error-in-variable estimators are considered.

4.4.1 Models based on Weighted Least Square Estimator

Initially, we estimate the model using the Weighted Least Square Estimator (WLSE) discussed in the previous section. Models based on other panel data estimators will then be investigated, for various reasons to be discussed subsequently. In the WLSE model, all the variables are weighted by the square root of the number of the observations in the cohort. As the log linear transformation is done on the individual household data, variables with value of zero can not be transformed. As a result, it is not possible to estimate double log model (many households have zero cars) and only linear and semi-log models are considered. The dependent variable is always the average number of car owned or used by the household, while the explanatory variables in the semi-log model include the cohort average of log-transformed income and motoring costs.

The first set of models considered is Pooled Weighted Least Square. Four models of linear form have been considered, whose difference lies in the representation of household structure and location. More specifically, Model 1 includes proportion of households in different household types (variable 'HH2' to 'HH8') and area types (variable 'Area2' to 'Area5') in each cohort[9]; Model 2 includes average number of children, working persons and household size (variable 'Child', 'Worker' and 'HHSize'), and rather than using five-way categorization of locations, it includes proportions of those living in metropolitan areas and least populated rural areas

[9] HH1 and Area1 are omitted, so the model evaluates the difference against the household in those categories.

(variable 'Met' and 'Rural'); Model 3 includes the proportion of household types and the proportion of household living in "Met" and "Rural"; Model 4 includes average number of children, working persons and household size as well as proportion of four area types. Besides the household structure and location variable, a constant term, average real disposable income ('Inc'), real car purchase price index ('Price'), real car running cost index ('RunCst'), average age of household head and its square ('Age' and 'AgSq') are all included in these models.

Similarly, twelve models of semi-log form have been investigated. These twelve models can be divided into three "blocks". Each block has the same household structure and location variables as the linear models, but has different logarithmic transformed variables. The first four models include the average of log disposable income (variable 'ALnInc') and all other variables remain in linear form; the second four models include the average of log disposable income, as well as log of car purchasing price and running costs ('ALnInc', 'LnPrice' and 'LnRunCst'); the third four models include not only the log transformed income and price variables, but also log of 'Age' and 'AgSq'.

Durbin-Watson autocorrelation test, White Heteroskedasticity test, and RESET specification test have been conducted for each model to identify model mis-specification. Adjusted R Square is used as an indicator to determine the model fit, while the sign and significance of the regression coefficients are also taken into account. Based on these criteria, the model with best fit is the semi-log model (dependent variable $y = A_{ct}$) with the following explanatory variables:

- Proportion of households as each of the seven household types[10];
- Proportion of households living in Metropolitan area including Greater London and the least populated rural area;
- Average of log household income and log car price and running costs;
- Second polynomial of average age of household head.

[10] Models based on alternative specification consistently produce a negative and significant coefficient for the number of children in the household, which is opposite to expectation. This is likely to be caused by high correlation between the number of children and household size.

The same search procedure is subsequently applied to the Fixed Effect Model. Among the sixteen models tested, the semi-log model with the same specification as above has the best fit. Table 4.2 reports the results of the Pooled WLS model and the Fixed Effect model.

Table 4-2 Regression results of Pooled WLS model and Fixed Effect Model

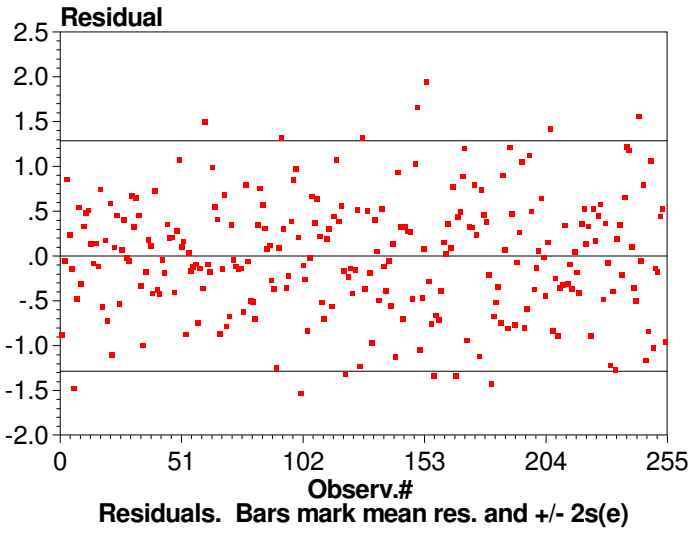
	Pooled WLS		Fixed Effect	
	Coeff	**t-ratio**	**Coeff**	**t-ratio**
Constant	0.0902	0.16		
ALNINC	0.3175	26.03	0.3050	24.99
HH2	-0.5031	-3.04	0.0242	0.14
HH3	-0.8278	-4.05	-0.1652	-0.79
HH4	0.1169	0.79	0.3795	2.90
HH5	0.0250	0.14	0.3035	1.86
HH6	0.2297	1.60	0.3090	2.22
HH7	0.6940	3.83	0.9693	5.87
HH8	0.4019	2.44	0.5440	3.52
MET	-0.3401	-3.22	-0.1687	-1.80
RURAL	0.2208	1.95	0.1415	1.49
LNPRICE	-0.1675	-5.57	-0.2098	-7.35
LNRUNCST	-0.0654	-2.16	-0.0676	-2.57
AGE	0.0170	4.95	0.0243	6.95
AGSQ	-0.0002	-4.92	-0.0003	-6.39
C1			0.8745	0.90
C2			0.7840	0.80
C3			-0.0015	0.00
C4			-0.6476	-0.66
C5			-1.5043	-1.50
C6			-1.8171	-2.09
C7			-1.3516	-1.77
C8			-1.0400	-1.47
C9			-0.2813	-0.43
C10			-0.3390	-0.53
C11			-1.0366	-1.83
C12			-1.6896	-2.87
C13			-2.1355	-3.31
C14			-2.3198	-3.19
C15			-2.9859	-3.95
C16			-4.2379	-4.83
Adjusted R^2	0.992		0.995	
SSE	166.33		104.92	
Log Likelihood	-306.62		-248.12	
DW stat	1.39		1.98	
F-stat of White Test	2.86		3.06	
t-stat of RESET Test	2.58		0.41	

Note: $y = A_{ct}$

Both models have high adjusted R square of over 0.99, suggesting most of the car ownership difference across cohorts can be captured by the explanatory variables.

While there is evidence of misspecification for the Pooled WLS model suggested by the RESET test, it seems that the Fixed Effect model is the appropriate specification. The Likelihood Ratio Test produces a Chi-square statistic of 117.0, which rejects the hypothesis of no cohort fixed effect at 1% level. For the Fixed Effect model Durbin-Watson Statistics does not reject the hypothesis of no auto-correlation; the F-Statistic of White Test is significant at 5% level, suggesting the possibility of heteroscedasticity or some form of mis-specification. Figure 4.1 shows the regression residual plot of the Fixed Effect model as produced by Limdep. Note that X axis shows the ID of cohort observations arranged by year for each cohort. The two middle horizontal bars mark zero plus and minus two times the estimated standard deviation of the residuals.

Figure 4-1 Regression Residual Plot of the Fixed Effect Model



Note: X-axis is ordered by year for each cohort

Based on the Fixed Effect model, the average log income variable is highly significant, and the coefficient of 0.305 suggests that a 1% increase in household income would lead to an increase of 0.003 cars per household. Note that the semi-log functional form implies that the impacts of income increase on car ownership gets smaller as the income level gets higher. The implied income elasticity at median car ownership level is 0.33. The log car purchasing price and log running costs variables are also significant at 1% level, and their coefficients imply that a 1% decrease of purchasing price and running costs would lead to increased ownership level of 0.002 and 0.0007 cars per household respectively. The implied purchase price and running cost elasticities are -0.23 and -0.07 respectively.

The household structure also has significant impacts on car ownership, as four out of seven "proportion of household type" variables are significant at 5% level. Not surprisingly, the biggest effects lie in proportion of "large" household, i.e. household with three or more adults (variable HH7 and HH8). A 1% increase of household type 7 (three or more adults, no children) within a cohort would imply an increase of 0.01 cars per household in that cohort. Similarly, a 1% increase of household type 8 (three or more adults, with children) would imply an increase of 0.005 cars per household[11].

Regarding the household location, while the coefficient of the RURAL variable (proportion of household living in least populated rural areas) is not significant, the proportion of household living in metropolitan areas has a significant impact on the car ownership of a cohort (at 10% level). The regression coefficient implies that when the proportion of household living in metropolitan area increases by 1%, the average number of cars per household decreases by 0.0017 for that cohort[12].

Both the average age of household head and its square are highly significant. Their coefficients are positive and negative respectively, suggesting the car ownership increases with the age of household head but at a decreasing rate. The "turning point" (marginal impacts turn from positive to negative) implied by the regression coefficient is 40 year of age, which is lower than the maximum car ownership age of 48 as identified in the data chapter (Chapter 3, Section 3.4.1). The cohort fixed effects fluctuate with a general downward trend for younger cohorts, which is opposite to the fact that younger cohorts tend to have higher car ownership level when other things being equal. More worryingly, this result is consistently obtained in all estimated semi-log models including those not reported here. It is a first hint that the empirical data do not support the assumption of a linear economic relationship between car ownership and other explanatory variables at household level (note that this assumption underlies

[11] As the proportion of household type 1 (single working person household) is dropped from the regression, a 1% increase of household type 8 (or any other types) means the proportion of household type 1 is reduced by 1%.

[12] A 1% increase of households living in metropolitan areas implies a 1% drop of those living in other rural areas with population density more than 2.2 persons per hectare (base case).

the Weighted Least Square Estimator use here). This point will become more apparent in the next Chapter on dynamic models.

4.4.2 Models Based on Genuine Panel Data Estimators

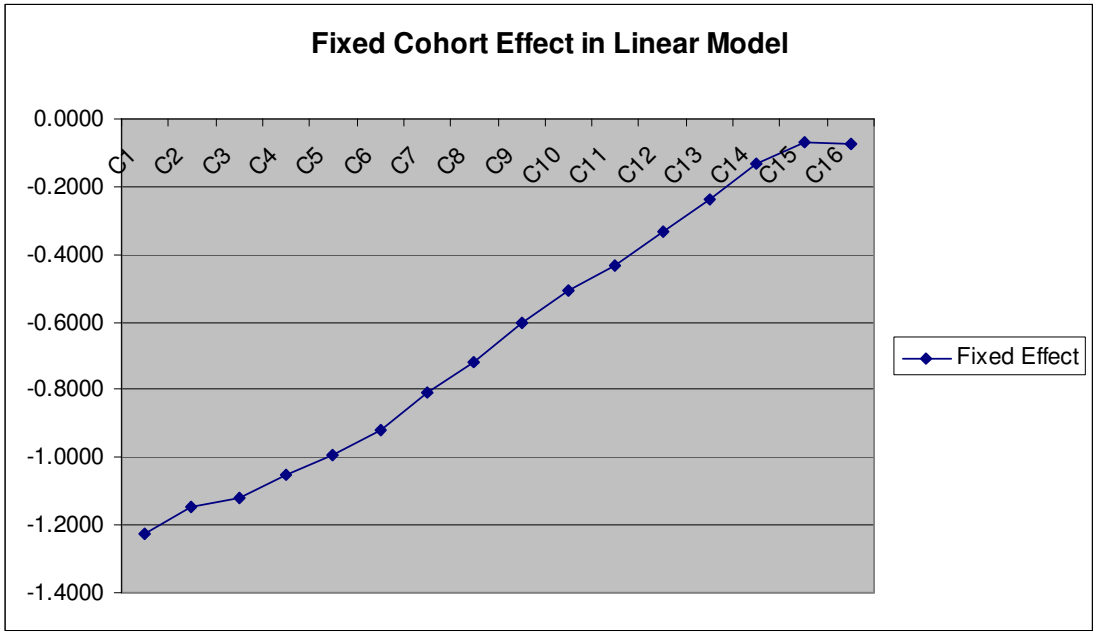
The Weighted Least Square estimator reported in the previous section is the Instrumental Variable estimator of individual household data. Such interpretation has two problems in the current study, both for analytical purpose and forecasting purpose. For analytical purpose, it treats the number of cars owned or used by each household as linear, i.e. the cause of car number increasing from zero to one is the same as that of any other unit increase. This is problematic as car ownership is in essence discrete choices made by household, with the decision to own the first car significantly different from the subsequent ones. For forecasting purpose, using a model based on individual household data will add difficult complications. It is only possible to predict the explanatory variables at aggregate cohort level, and linear transformation (log, square, etc.) can only be done for that average prediction (e.g. log of average household income). When the model is estimated based on the average of linear transformed data (e.g. average of log household income), aggregation bias would arise.

Given these problems, it is important to interpret the pseudo panel model from a different perspective. Recall that we only include cohorts with sufficiently large number of observations, and we believe the measurement error can be ignored in this case. This implied that the sample averages of each cohort are unbiased estimates of true cohort means, and as a result, the pseudo panel can be estimated as a genuine panel with each cohort being the unit of observation. In this case, any linear transformation of the data will be done at the cohort level and the panel data fixed effect and random effect models can be estimated using standard technique.

Systematic specification search procedure similar to the WLS estimator models has been carried out. Three functional forms have been tested for the Fixed Effect Model: linear, semi-log linear (of various explanatory variables) and double log linear form. The double log form implies a constant income and price elasticity (note that it can not be estimated using the WLS estimator due to zero car ownership for some households). For each functional form, the explanatory variables include average real household disposable income (or its log); household structure variables (number of children and

working person and household size; alternatively, proportion of household types); household location variables (proportion of household living in four areas; alternatively, proportion of household living in metropolitan and least populated rural area); index of real car purchasing price and real car running costs; and finally, second polynomial of average age of household head.

Figure 4-2 Fixed effects in the linear model



There is a prevalent feature across most fixed effect models examined here: linear trend in the fixed cohort effect. Figure 4.2 shows the fixed effect in the linear model with household structure variables being the proportion of seven household types and location variable being "Met" and "Rural". It shows the linear trend in the fixed effects of most cohorts except for the youngest one. It should be noted that there are high correlations between the fixed effects and some of them are not statistically significant.

Consequently, the restricted Fixed Effect Models have been estimated, with the sixteen cohort dummy variables replaced by a variable of cohort number (ID) and a dummy for the youngest cohort (cohort 16). It implies that the fixed effects across cohorts are linear except for the youngest cohort, which is similar to the "Generation Model" of Dargay and Vythoulkas (1999). For each of the unrestricted and restricted FE models, standard set of RESET misspecification test, Durbin-Watson autocorrelation test and White heteroscedasticity test has been conducted. The RESET test does not indicate

Table 4-3 Linear Model: Unrestricted and restricted Fixed Effect Model

	Linear Model				Semi Log Linear Model			
	Unrestricted		Restricted		Unrestricted		Restricted	
	Coeff	t-ratio	Coeff	t-ratio	Coeff	t-ratio	Coeff	t-ratio
Constant			-1.0872	-4.12			-1.7978	-2.41
INC	0.0006	3.80	0.0009	6.64				
LNINC					0.2163	4.82	0.2732	6.26
HH2	-0.1553	-0.95	-0.4209	-2.92	-0.1478	-0.91	-0.4569	-3.13
HH3	-0.7297	-3.47	-0.5625	-2.98	-0.6830	-3.28	-0.5814	-3.04
HH4	0.4722	3.54	0.1230	0.97	0.4321	3.29	0.0595	0.47
HH5	0.5961	3.76	0.2067	1.42	0.4827	3.00	0.0867	0.58
HH6	0.7194	5.20	0.3540	2.83	0.6187	4.37	0.2976	2.29
HH7	1.0214	5.59	0.5644	3.27	0.9863	5.44	0.5747	3.29
HH8	0.9223	5.30	0.4725	2.82	0.8430	4.86	0.4872	2.87
MET	-0.2434	-2.37	-0.2537	-2.72	-0.2116	-2.08	-0.2496	-2.63
RURAL	0.1494	1.34	0.1548	1.53	0.1201	1.10	0.1699	1.66
PRICE	-0.0020	-2.39	-0.0008	-0.91				
RUNCST	-0.0023	-4.34	-0.0012	-1.99				
LNPRICE					-0.1256	-1.53	-0.0417	-0.46
LNRUNCST					-0.1910	-3.46	-0.1202	-1.89
AGE	0.0455	13.43	0.0381	12.28	0.0439	12.91	0.0375	11.74
AGSQ	-0.0003	-9.06	-0.0002	-8.31	-0.0003	-8.66	-0.0002	-7.64
COHORT			0.0706	7.53			0.0736	8.01
C1	-1.2248	-5.06			-1.1537	-1.67		
C2	-1.1449	-4.87			-1.0729	-1.57		
C3	-1.1175	-4.85			-1.0443	-1.53		
C4	-1.0498	-4.65			-0.9810	-1.44		
C5	-0.9947	-4.49			-0.9290	-1.37		
C6	-0.9192	-4.25			-0.8563	-1.27		
C7	-0.8102	-3.83			-0.7479	-1.11		
C8	-0.7164	-3.47			-0.6525	-0.97		
C9	-0.5999	-3.00			-0.5312	-0.79		
C10	-0.5089	-2.61			-0.4390	-0.65		
C11	-0.4315	-2.28			-0.3654	-0.54		
C12	-0.3352	-1.83			-0.2733	-0.41		
C13	-0.2375	-1.34			-0.1813	-0.27		
C14	-0.1305	-0.76			-0.0806	-0.12		
C15	-0.0688	-0.41			-0.0183	-0.03		
C16	-0.0748	-0.46	-0.1341	-3.90	-0.0258	-0.04	-0.1354	-3.94
RHO			0.3511	5.97			0.3307	5.57
Adjusted R^2	0.986		0.983		0.986		0.985	
SSE	0.31		0.340		0.30		0.26	
Log Likelihood	491.70		479.59		495.27		476.42	
RESET (t-Stat)	-.066		0.74		-0.15		2.69	

Note: $y = A_{ct}$ in both models

misspecification for most of the un-restricted models. While the same test suggests misspecification for the all restricted semi-log and double-log models, it does not reject the null hypothesis of no misspecification for the linear models at any statistically significant level.

For the unrestricted FE models, the Durbin-Watson Test and White Test do not clearly reject the hypothesis of no autocorrelation and homoscedasticity. For the restricted models, these hypotheses are evidently rejected by the same tests, so they have been re-estimated using Feasible Generalized Least Square with AR1. Table 4.3 compares the results of unrestricted and restricted fixed effect models of best fit, both with linear and semi-log linear functional form.

The regression coefficients are quite similar for all models (Except for those with log linear transformation). The adjusted R Square varies between 0.983 and 0.986, which indicate good level of fit. In terms of model selection, the first step is to compare the unrestricted and restricted models. The likelihood ratio test based on the linear models produces a Chi Square statistic of 24.2, and with 14 degree of freedom, the hypothesis of no loss of fit is rejected at 5% level. The same test based on semi-log model produced a Chi Square statistic of 37.7, which rejects the hypothesis at 1% level. In both cases, the unrestricted fixed effect models are chosen as favoured.

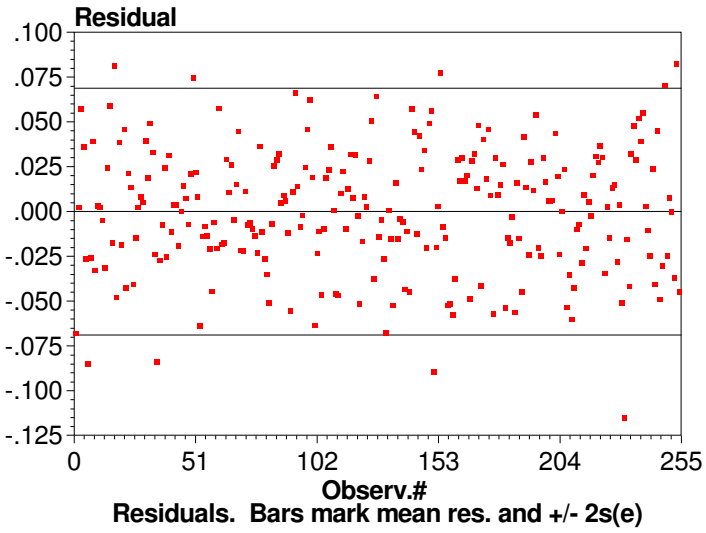
The differences between the linear and semi-log models are more subtle. For the linear model, the coefficients of "Price" and "RunCst" are higher (in absolute term) than the coefficient of "Inc". Based on median income and car ownership level, the implied income elasticity is 0.20, the implied price elasticity is -0.21 and the implied running costs elasticity is -0.25. As previous research suggests that income elasticity is higher than price/running costs elasticity, such results are troublesome. Consequently, the (unrestricted) semi-log model is chosen as the preferred model, whose residual plot in Figure 4.3 does not show any apparent misspecification.

Based on the preferred model, the log income coefficient suggests that a 1% increase of real household disposable income will increase the car ownership by 0.0021 cars per household. On the other hand, a 1% decrease in car purchase price will increase the average car ownership level by 0.0013 cars[13]; a similar decrease in running costs will lead to a rise in car ownership level by 0.0019 cars. Table 4.4 shows the income and

[13] Note that the coefficients for purchasing price are not significant, so there should be caution in interpretation of this result.

price elasticities at the low (10 percentile), median and high (90 percentile) car ownership level.

Figure 4-3 Residual Plot of unrestricted Fixed Effect Model



Note: X-axis is ordered by year for each cohort

Table 4-4 Income and Price Elasticity (based on Semi log, unrestricted FE model)

Car Ownership Level	Income Elasticity	Purchase Price Elasticity	Running Cost Elasticity
Low (0.42 cars/HH)	0.52	-0.30	-0.45
Median (0.92 cars/HH)	0.24	-0.14	-0.21
High (1.25 cars/HH)	0.17	-0.10	-0.15

Regarding the household structure, the higher is the proportion of "large" household (household with two or three adults) in a cohort, the higher is the car ownership level. The impacts of number of working person and children is best illustrated by the two adult household (type 4, 5 and 6). When the proportion of household type 4 (two adults, neither in work) increases by 1%[14], the average number of car increases by 0.0043 per household in the cohort; when the proportion of household type 5 (two adults with working person but no children) increases by 1%, the average car number increases by 0.0048; when the proportion of household type 6 (two adults with working person and children) increases by 1%, the average car number increases by 0.0062. Regarding the household location variables, only "Met" is significant, whose coefficient suggests that

[14] It implied the proportion of household type 1 decreases by 1%; same for other household type changes.

a 1% increase of Metropolitan household in a cohort results in a decrease of 0.0021 cars per household in that cohort.

Besides the Fixed Effect model, a set of Random Effect Models has been estimated. Assuming that the cohort effects are strictly uncorrelated with the explanatory variables, Random Effect models treat the cohort specific constant terms as randomly distributed across cohorts ($\lambda+v_i$ rather than λ_i). With fewer parameters to estimate, such specification increases the degree of freedom. However, in the context of pseudo panel, the assumption of no correlation between the cohort effects and the regressors is likely to be violated, and the RESET test reveals misspecification in most of the random effect models tested. The presence of correlations seems to have been confirmed by the Hausman's Test, as the Chi Square statistics are large and statistically significant for all the models tested. For example, for the linear model reported in Table 4.3, its corresponding Random Effect model has a Hausman's Test statistic of 70.9, which strongly favours the Fixed Effect model.

Finally, we investigated the issue of heterogeneity by estimating cohort specific models. The oldest and two youngest cohorts have to be dropped due to small number of observations and the resulting dataset contains two cohorts with 14 observations (C2 and C14) and eleven cohorts with 19 observations (C3 to C13). To save the regression degrees of freedom, average household characteristic variables (number of children, working person and household size) rather than proportion of household types are used. The initial modelling group tested contains eleven explanatory variables including the constant term in each model, separately estimated using OLS. Three variables that are not significant for almost all cohorts are subsequently dropped, which leaves 6 error degree of freedom for the models with 14 observations and 11 error degree of freedom for those with 19 observations. Different groups with linear and semi-log linear functional form have been estimated and the results are very similar (in terms of coefficients of the common variables and R square). Table 4.5 reports the results of the linear models for each of the 13 cohorts.

For the ease of comprehension, only the coefficients that are significant at 10% level are reported; however, the descriptive statistics are calculated using all 13 coefficients. The last two column of Table 4.5 refer to the standard error of regression and R Square

Table 4-5 Cohort Specific Regression Results (Linear form; 13 cohorts)

	Inc	Child	HHSize	Met	Price	Age	AgSq	Const.	SSE	R^2
C2	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.049	0.876
C3	n.s.	n.s.	n.s.	-0.7390	-0.0063	0.1961	-0.0014	-5.9712	0.032	0.961
C4	0.0010	n.s.	0.3331	-0.4776	n.s.	0.1561	-0.0011	-5.1904	0.027	0.967
C5	n.s.	-1.4233	0.9822	n.s.	n.s.	0.1657	-0.0011	-7.0443	0.020	0.984
C6	0.0019	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.037	0.929
C7	n.s.	n.s.	n.s.	n.s.	-0.0084	n.s.	n.s.	6.8813	0.022	0.974
C8	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.038	0.871
C9	n.s.	-0.6406	0.8139	n.s.	n.s.	n.s.	n.s.	n.s.	0.046	0.920
C10	n.s.	-0.4308	0.4989	1.1212	n.s.	n.s.	n.s.	-1.5895	0.031	0.975
C11	0.0012	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.023	0.987
C12	n.s.	n.s.	n.s.	n.s.	n.s.	0.1397	-0.0018	-3.1804	0.029	0.981
C13	0.0014	-0.6733	0.6033	n.s.	-0.0028	0.1353	-0.0019	-2.4892	0.014	0.998
C14	n.s.	n.s.	n.s.	n.s.	-0.0065	0.2919	-0.0051	-2.7124	0.024	0.995
Mean	0.0006	-0.3446	0.3864	-0.1598	-0.0029	0.1368	-0.0013	-3.6185	0.030	0.955
St Dev	0.0006	0.5673	0.3443	0.4816	0.0030	0.1271	0.0014	4.9290	0.010	0.043
Max	0.0019	0.5705	0.9822	1.1212	0.0006	0.3937	0.0005	6.8813	0.049	0.998
Min	0.0000	-1.4233	-0.2079	-0.7390	-0.0084	-0.0968	-0.0051	-14.8316	0.014	0.871

respectively. The R Square varies from 0.871 to 0.998 among the 13 models, which indicates the same model specification has different explanatory powers in describing the data for different cohorts. For the two cohort models with R Square lower than 0.9 (Cohort 2 and Cohort 8), none of the explanatory variables are significant at 10% level.

The pairwise tests on coefficients of the "Income" variable show no statistically significant difference between the 4 coefficients reported in Table 4.4. However, such tests show that some coefficient pairs are different at 5% level for other explanatory variables, which are indication of parameter heterogeneity. The hypothesis of homogeneity is also rejected by the Likelihood Ratio test of parameter homogeneity, which is based on a Random Parameter Model we subsequently estimated.

The above discussion indicates that the assumption of parameter homogeneity for the earlier models might not be appropriate. However, it also highlights the practical constraints faced by the heterogeneous model in terms of degree of freedom lost. Table 4.5 shows that a majority of the cohort specific regression coefficients are not significant at 10% level. Furthermore, several cohorts have too few observations to be included in the cohort specific regression in the first place. These constraints will limit the practical usefulness of the heterogeneous model.

4.5 Conclusion

This chapter investigates the theoretical and empirical aspects of the linear static pseudo panel model. As the pseudo panel is a synthetic panel constructed from individual data, the first issue we address is the relationship between the estimator based on pseudo panel and that based on individual data. It has been shown that Weighted Least Square estimator using cohort means is equivalent to the Instrumental Variable estimator using individual data, using cohort dummy variables as instruments. However, this relationship is based on the assumption that there is a linear relationship between the dependant variables and the explanatory variables for the individual household. For car ownership model, this assumption is hard to defend theoretically and its appropriateness has to be tested with empirical models. If Weighted Least Square estimator is to be used, it requires any linear transformation of data to be performed on individuals before aggregation, and each observation in the pseudo panel to be weighted by the square root of the sample number in the cohort.

A number of papers have addressed the problem of measurement error for the pseudo panel model. Section 2 of this chapter reviews the various Error-in-Variable Estimators (EVE) proposed in the literature, including that of Deaton (1985), which is unbiased when $T \to \infty$; that of Verbeek and Nijman (1993), which is unbiased when $C \to \infty$; and that of Devereux (2003), which is approximately unbiased to the order $1/(CT)$. The following section attempts to address the conditions required to ignore the measurement error problem. It has been shown that the way cohorts are constructed has direct implication on the bias of the within estimator if pseudo penal is to be estimated as genuine panel. The cohort should be defined in a way such that the population cohort means of the variables concerned vary as much as possible over time. Furthermore, the sample number in each cohort has to be sufficiently large to minimize sampling errors and measurement errors. These conditions appear to be met by our pseudo panel dataset constructed from the Family Expenditure Survey, which justify us to ignore the problem of measurement error in the following empirical work.

We first estimate the car ownership model using Weighted Least Square Estimator of pseudo panel. Different formulations of household structure variables and household location variables have been systematically tested, and the results shows that they are

best represented by the split of 8 household types and aggregated location types respectively. Both the Likelihood Ratio Test and RESET test favour Fixed Effect model to Pooled OLS model. Judged by the statistic of adjusted R Square, it appears that the semi-log models (explanatory variables include the cohort average of log-transformed income and price variables) have the best goodness of fit. However, all semi-log models have the wrong trend of fixed effects across cohorts, implying older cohorts have higher tendency of car ownership when other things being equal. This unsatisfactory result could be taken as the first hint that the empirical data do not support the use of Weighted Least Square Estimator here.

From a theoretical point of view, the Weighted Least Square Estimator is based on the assumption of linear economic relationship at individual household level, which would be inappropriate for car ownership due to its discrete nature. Furthermore, the use of WLSE requires any linear transformation of variables to be done at individual level, which could not be replicated for the forecasted value. These concerns prompt us to consider models based on alternative estimators. If we believe the sample averages of each cohort are unbiased estimates of the true cohort means, the pseudo panel can be treated as genuine panel. In this case, various panel data estimators including fixed effect, random effect and heterogeneous models have been investigated.

Hausman's Test and RESET misspecification test both reject the Random effect model as the preferred model. Regarding the Fixed Effect model, there is a strong linear trend in the cohort specific constants; however, the likelihood ratio test reject restricted FE model as the preferred model due to significant loss of fit. Based on the semi-log unrestricted FE model, the implied income elasticity is 0.24, purchase price elasticity is -0.14 and running cost elasticity is -0.21.

Despite a rather comprehensive investigation of the pseudo panel car ownership model from both theoretical and empirical perspectives, one important aspect is not covered in this chapter: the role of dynamics. The dynamic models will be considered in the following chapter.

Chapter 5 Linear Dynamic Model

One of the key motivations of using pseudo panel rather than cross sectional data is to capture the dynamic relationships in individual data. After the discussion of consistent estimation of static model in the previous chapter, we focus our attention on the dynamic model. More specifically, we consider the first order autoregressive model with exogenous variables, which is prevalent in the literature. In this chapter, we outline a common set of assumptions and rank conditions, and then consider various estimators that are consistent under different asymptotics. We also consider their empirical implication including practical implementation, data requirement and complexity of computation. We present an estimator that is computationally attractive and consistent when the number of observations in each cohort is large. However, we will not cover topics such as estimation of asymptotic covariance matrices or efficiency of the proposed estimators, which are considered less important for the main purpose of this study—forecasting. After discussing consistent estimation of pseudo panel model, the second part of this chapter will present the empirical results of dynamic car ownership models using the preferred estimator. Systematic specification search has been carried out to determine the model with the best fit. Various statistical tests are used to aid model selection.

5.1 Consistent estimator of dynamic pseudo panel model

We first consider a dynamic model for each individual i sampled in year t, whose explanatory variables include a lagged dependent variable and a vector of exogenous variables. Note the time-varying exogenous variables as $x_{i(t),t}$ [15] and the time-invariant exogenous variables as $z_{i(t)}$:

$$y_{i(t),t} = \alpha y_{i(t),t-1} + x'_{i(t),t}\beta + z'_{i(t)}\lambda + \varepsilon_{i(t),t} \qquad t = 2,\ldots,T \qquad (1)$$

with the following assumptions:

$$\varepsilon_{i(t),t} \sim \text{iid}\,(0, \sigma_\varepsilon^2) \qquad (2)$$

$$E\{\varepsilon_{i(t),t} z_{i(t)}\} = 0 \qquad (3)$$

and $$E\{\varepsilon_{i(t),t} x_{i(t),t}\} = 0 \qquad (4)$$

[15] The notation of $x_{i(t),t}$ reflects the fact that the ith individual sampled in year t is different from that sampled in other year, e.g. individual $i(t)$ is different from individual $i(t+1)$.

The iid assumption of the error term is common in the pseudo panel literature, with Moffitt (1993), Collado (1997) and most of McKenzie (2004) all based on such assumptions [16]. Conditions (3) and (4) are standard exogeneity assumption of independent variables, which are easy to defend for individual observations in the repeated cross sections. Note that with repeated cross sections data, the lagged dependant variable $y_{i(t),t\text{-}1}$ is not observable. As a result, model (1) can not be directly estimated.

As identified in the literature, the dynamic pseudo panel models can be consistently estimated using two approaches: first by grouping individuals in the survey data into cohorts and treat the cohort averages as error-ridden estimation of the observation in the synthetic panel, which also provides an error-ridden estimation of the lagged dependent variable at cohort level; second is to directly estimate such model from cross sectional data using instrumental variable techniques, typically by replacing the lagged dependent variable by a predicted value from an auxiliary regression. As we have shown in the previous chapter, taking group averages is equivalent to using cohort dummy variables as instruments and applying the IV estimator on individual data. This suggests that there are no fundamental differences between the two approaches. Nevertheless, we present them separately in this chapter with emphasis on the first approach, i.e. cohort dummy IV estimator.

5.1.1 Cohort Dummy IV estimator

As $z_{i(t)}$ represent time-invariant characteristics, we would be able to divide the population into a number of mutually exclusive cohorts based on $z_{i(t)}$. In such case, $z_{i(t)}$ can be expressed as a vector of cohort dummy variables (z_c). Instrumenting $x_{i(t),t}$ using z_c would lead to estimators based on cohort average $\bar{x}_{ct}$ (see previous chapter for details). Take average of n_{ct} individuals in cohort c and weight it by the square root of n_{ct}, equation (1) becomes:

$$\bar{y}_{c(t),t} = \bar{\alpha} \cdot \bar{y}_{c(t-1),t-1} + \bar{x}'_{c(t),t}\bar{\beta} + z'_c\bar{\lambda} + \bar{\varepsilon}_{c(t),t} \qquad c = 1,\ldots,C;\ t = 2,\ldots,T \tag{5}$$

[16] However, we'll show that when $n_{ct} \to \infty$ this assumption can be relaxed while the estimators concerned remain consistent.

where $\bar{x}'_{c(t),t} = (1/n_{ct})\sum_{i=1}^{n_{ct}} x'_{i(t),t}$, denoting the sample mean of x over the individuals in cohort c observed in year t. Note that $\bar{y}_{c(t-1),t-1}$ denote the mean of y for individuals in cohort c sampled in year t-1, rather than the mean of lagged y for those sampled in year t, which is unobservable. As z_c is a vector of cohort dummy variables, $\bar{\lambda}$ represents the cohort fixed effects.

Equation (5) corresponds to an unobserved version based on true population cohort mean:

$$y^*_{ct} = \alpha y^*_{c,t-1} + x^*_{ct}{}'\beta + z'_c\lambda + \varepsilon^*_{ct} \qquad c = 1,\ldots,C;\ t = 2,\ldots,T \qquad (6)$$

Comparing (5) and (6), we can see that $\bar{x}'_{c(t),t}$ and $\bar{y}_{c(t),t}$ are the sample estimate of $x^*_{ct}{}'$ and y^*_{ct} but with measurement error. Similar to static model discussed in the previous chapter, we assume the measurement errors follow independent identical distribution:

$$\begin{pmatrix} \bar{y}_{c(t),t} - y^*_{ct} \\ \bar{x}_{c(t),t} - x^*_{ct} \end{pmatrix} \sim iid\left(0, \begin{pmatrix} \sigma^2_\varsigma & \sigma'_{\varsigma\eta} \\ \sigma_{\varsigma\eta} & \Sigma_\eta \end{pmatrix}\right) \qquad \forall c,t \qquad (7)$$

where σ^2_ς, $\sigma_{\varsigma\eta}$ and Σ_η can be consistently estimated by individual data. Note that from (7) it follows that $\bar{y}_{c(t),t} - y^*_{ct}$ and $\bar{y}_{c(t-1),t-1} - y^*_{c,t-1}$ are independent and have the same distribution.

Finally, for (5) to be identified, the following rank condition has to be satisfied:

$$\tilde{U} = \left(\bar{Y}_{-1} - \bar{\bar{y}}_{c,-1} \quad \bar{X} - \bar{\bar{x}}'_c\right) \text{ has full column rank K+1.} \qquad (8)$$

Note $\tilde{U}$ is the matrix of x and lagged y at cohort level deviating from overall cohort means. In particular:

$\bar{Y}_{-1}$ is a $C*(T-1)\times 1$ vector whose element is $\bar{y}_{c(t-1),t-1}$, $c = 1,\ldots,C;\ t = 2,\ldots,T$.

$\bar{\bar{y}}_{c,-1} = \dfrac{1}{T-1}\sum_{t=2}^{T}\bar{y}_{c(t-1),t-1}$, $c = 1,\ldots,C$. It denotes the average of sample cohort means of y over year 1 to $T-1$.

$\bar{X}$ is a $C*(T-1)\times K$ matrix with row $\bar{x}'_{c(t),t}$, $c = 1,\ldots,C;\ t = 2,\ldots,T$.

$\bar{\bar{x}}'_c$ is a $1\times K$ vector: $\bar{\bar{x}}'_c = \dfrac{1}{T-1}\sum_{t=2}^{T}\bar{x}'_{c(t),t}$, $c = 1,\ldots,C$. It denotes the average of sample cohort means of x over year 2 to T.

Such rank condition is a standard identification condition. It requires that cohort means of *y* and *x* should not exhibit perfect collinearity and vary over time. It also implies that there should be at least three cross sections for the model to be identified. If (8) is not satisfied, the estimation of (5) will break down.

With repeated cross sectional data, it is important to establish the properties for different estimators under alternative asymptotics. In the following sections, we present the various estimator identified in the literature as being consistent under the asymptotic of $T \to \infty$, $C \to \infty$ and $n_{ct} \to \infty$ respectively.

5.1.1.1 Error Corrected Within-Group Estimator

Following Nickell (1981) on consistent estimation of dynamic model based on true panel data, Collado (1997) proposed a within-group estimator taking into account measurement error variances, which is consistent when $T \to \infty$. Note that the error term $\bar{\varepsilon}_{c(t),t}$ in (5) contains the measurement error of x^*_{ct} and y^*_{ct}, and can be decomposed into:

$$\bar{\varepsilon}_{c(t),t} = \varepsilon^*_{ct} + (\bar{y}_{c(t),t} - y^*_{ct}) - \alpha(\bar{y}_{c(t-1),t-1} - y^*_{c,t-1}) - (\bar{x}'_{c(t),t} - x^{*}_{ct}{}')\beta \qquad t = 2,\ldots,T \tag{9}$$

Rewrite equation (5) in the form of deviation from cohort means to eliminate the cohort fixed effects:

$$\tilde{y}_{c(t),t} = \hat{\alpha}\tilde{y}_{c(t-1),t-1} + \tilde{x}'_{c(t),t}\hat{\beta} + \tilde{\varepsilon}_{c(t),t} \tag{10}$$

Combing (7) and (9), Collado showed that under asymptotic of $T \to \infty$, the explanatory variables in (10) are correlated with the error term only through the measurement error in the following way:

$$E\left[\begin{pmatrix} \tilde{y}_{c(t-1),t-1} \\ \tilde{x}_{c(t),t} \end{pmatrix}\tilde{\varepsilon}_{c(t),t}\right] \to \begin{bmatrix} -\alpha\sigma^2_{\varsigma} \\ \sigma_{\varsigma\eta} - \Sigma_{\eta}\beta \end{bmatrix} \qquad \text{as } T \to \infty \tag{11}$$

This leads to the measurement error corrected within-group estimator (WGC) of:

$$\begin{pmatrix} \hat{\alpha}_{wge} \\ \hat{\beta}_{wge} \end{pmatrix} = \left[\frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{t=2}^{T}\begin{pmatrix} \tilde{y}^2_{c(t-1),t-1} & \tilde{y}_{c(t-1),t-1}\tilde{x}'_{c(t),t} \\ \tilde{x}_{c(t),t}\tilde{y}_{c(t-1),t-1} & \tilde{x}_{c(t),t}\tilde{x}'_{c(t),t} \end{pmatrix} - \begin{pmatrix} \sigma^2_{\varsigma} & 0 \\ 0 & \Sigma_{\eta} \end{pmatrix}\right]^{-1}$$

$$* \left[\frac{1}{C(T-1)} \sum_{c=1}^{C} \sum_{t=2}^{T} \begin{pmatrix} \tilde{y}_{c(t-1),t-1} \tilde{y}_{c(t),t} \\ \tilde{x}_{c(t),t} \tilde{y}_{c(t),t} \end{pmatrix} - \begin{pmatrix} 0 \\ \sigma_{\varsigma\eta} \end{pmatrix} \right] \tag{12}$$

It is expected the error corrected within-group estimator would have limited use in the empirical work, as it is unlikely that a large number of repeated cross sections are available in reality so that the condition of $T \to \infty$ is satisfied. Furthermore, the requirement to estimate the variance and covariance of measurement error significantly increases the computation complexity, thus further constrains its scope of application.

5.1.1.2 Error Corrected GMM Estimator

As $T \to \infty$ can be an unlikely situation in applied work, Collado presents an alternative estimator that is consistent for finite T when the number of cohort C is large (Collado, 1997).

Following standard dynamic panel model IV procedures, the cohort fixed effects can be eliminated by first differencing. Rewrite equation (5) in the form of first differences:

$$\Delta \bar{y}_{c(t),t} = \hat{\alpha} \Delta \bar{y}_{c(t-1),t-1} + \Delta \bar{x}'_{c(t),t} \hat{\beta} + \Delta \varepsilon_{c(t),t} \quad c = 1,\ldots,C;\ t = 3,\ldots,\mathrm{T} \tag{13}$$

where $\Delta \bar{y}_{c(t),t} = \bar{y}_{c(t),t} - \bar{y}_{c(t-1),t-1}$, which is a sample estimate of $y^*_{ct} - y^*_{c,t-1}$ but with measure error; $\Delta \bar{y}_{c(t-1),t-1}$ and $\Delta \bar{x}'_{c(t),t}$ are defined similarly. Note that the explanatory variables in (13) are correlated with the error terms $\Delta \varepsilon_{c(t),t}$ through measurement errors.

Assuming the error terms follow iid (assumption 2) and the explanatory variables are strictly exogenous (assumption 3 and 4), Collado proposed the following matrix of instruments[17] following the genuine panel case of Arellano and Bond (1991):

$$V_c = \begin{bmatrix} \bar{y}_{c1}, \bar{x}'_{c1}, \ldots, \bar{x}'_{cT} & 0 & \ldots & 0 \\ 0 & \bar{y}_{c1}, \bar{y}_{c2}, \bar{x}'_{c1}, \ldots, \bar{x}'_{cT} & \ldots & 0 \\ & \ldots & & \\ 0 & \ldots & 0 & \bar{y}_{c1}, \ldots, \bar{y}_{c,T-2}, \bar{x}'_{c1}, \ldots, \bar{x}'_{cT} \end{bmatrix}$$

[17] Collado note such matrix as Z_c. Here we use V_c to avoid confusing with the cohort dummy instruments $z_{i(t),t}$. Here, $\bar{y}_{c(t),t}$ is simply noted as $\bar{y}_{ct}$; similarly, $\bar{x}'_{c(t),t}$ is noted as $\bar{x}'_{ct}$.

V_c is correlated with the error terms only due to the measurement errors and the resulting moment conditions are given by: $E\{V_c'\Delta\varepsilon_c\} = \Lambda\gamma + \kappa$,

Where $\Delta\varepsilon_c = \begin{bmatrix} \Delta\varepsilon_{c(3),3} \\ ... \\ \Delta\varepsilon_{c(T),T} \end{bmatrix}$, a $(T-2)\times 1$ vector; $\gamma = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, a $(K+1)\times 1$ vector. Λ and κ depend on the covariance of measurement errors.

The measurement error corrected GMM estimator (GMMC) of γ is obtained by minimizing:

$$\sum_{c=1}^{C}(V_c'\Delta\varepsilon_c - \Lambda\gamma - \kappa)'S_c\sum_{c=1}^{C}(V_c'\Delta\varepsilon_c - \Lambda\gamma - \kappa)$$

and was derived by Collado as:

$$\hat{\gamma} = [\sum(\Delta W_c'V_c + \Lambda')S_c\sum(V_c'\Delta W_c + \Lambda)]^{-1}\cdot[\sum(\Delta W_c'V_c + \Lambda')S_c\sum(V_c'\Delta y_c - \kappa)] \quad (14)$$

where $W_c = \begin{bmatrix} \bar{y}_{c(2),2} & \bar{x}'_{c(3),3} \\ ... & ... \\ \bar{y}_{c(T-1),T-1} & \bar{x}'_{c(T),T} \end{bmatrix}$, $y_c = \begin{bmatrix} \bar{y}_{c(3),3} \\ ... \\ \bar{y}_{c(T),T} \end{bmatrix}$ and the weighting matrix S_c is a consistent estimator of the inverse of the covariance matrix of $V_c'\Delta\varepsilon_c$.

The error corrected GMM estimator is consistent for fixed T when $C \to \infty$. When the population is divided into a large number of cohorts based on various time-invariant characteristics, with large repeated cross sectional survey dataset the corresponding sample of each cohort should be available. In this case, GMMC provides a consistent estimation approach. However, GMMC also has drawbacks. It uses $\bar{y}_{c(t-s),t-s}$ for all s $(t-1 \geq s \geq 2)$ as instrument of $\Delta\bar{y}_{c(t),t}$. The resulting large number of instruments, while adding estimation efficiency, would also introduce finite-sample bias (McKenzie, 2004). It is also very complex computationally, given the need to estimate covariance matrix of the measurement errors and GMM weighting matrix S_c.

5.1.1.3 Within Group Estimator

Although the two error-corrected estimators by Collado are consistent with either a large number of time periods or a large number of cohorts, these are in fact not

common situations in empirical studies. More often, we have a limited number of cross sections, and each can be divided into a fixed number of cohorts. If each cross section contains a large number of observations, as is the case of many national surveys, we would have sufficiently large sample size in each cohort (n_{ct}). In this case, the most useful estimator would be the one that is consistent when $n_{ct} \to \infty$.

When $n_{ct} \to \infty$, it is obvious that the measurement errors (7) converge in probability to zero:

$$P\lim\begin{pmatrix} \bar{y}_{c(t),t} - y^*_{ct} \\ \bar{x}_{c(t),t} - x^*_{ct} \end{pmatrix} = 0 \qquad (15)$$

Combining (15) and the iid assumption of the measurement errors, it directly follows that the last three terms in the right hand side of (9) are all asymptotically zero. As a result, $\bar{\varepsilon}_{c(t),t}$ will have the same asymptotic mean as the true cohort error term ε^*_{ct}, whose mean is assumed to be zero as standard.

$$p\lim(\bar{\varepsilon}_{c(t),t}) = 0 \qquad (16)$$

From (16) and assumptions (2) to (4), we establish the moment conditions that will ensure the consistency of within-group estimator based on (5):

$$E\{\bar{\varepsilon}_{c(t),t}\bar{y}_{c(t-1),t-1}\} = 0 \qquad (17)$$

$$E\{\bar{\varepsilon}_{c(t),t}\bar{x}_{c(t),t}\} = 0 \qquad (18)$$

The rank condition of (8) is also required to ensure (5) is identified. Rewrite (5) in the form of deviation from cohort means to eliminate the fixed cohort effect, it results in the within-group estimator in its standard form:

$$\gamma_{within} = (\tilde{U}'\tilde{U})^{-1}\tilde{U}'\tilde{Y} \qquad (19)$$

where $\gamma = (\alpha, \beta')'$. $\tilde{U}$ is as defined in (8). $\tilde{Y} = \bar{Y} - \bar{\bar{y}}_c$, with $\bar{Y}$ being a $C*(T-1)\times 1$ vector whose element is $\bar{y}_{c(t),t}$, $c = 1,\ldots,C$; $t = 2,\ldots,T$; $\bar{\bar{y}}_c = \frac{1}{T-1}\sum_{t=2}^{T}\bar{y}_{c(t),t}$, $c = 1,\ldots,C$.

It should be noted that assumption (2) can be relaxed to allow for correlations between observations within the same cohort, as long as there is no autocorrelation over time. In this case, (17) to (18) also hold so the consistency of within estimator is retained.

The within group estimator is simple to implement in applied work and does not require very strong assumptions on the correlation structure of the error terms. It is consistent when the number of observations per cohort is sufficiently large, which is likely to be met with a real dataset. Ultimately, it is equivalent to the "Augmented IV Estimator" in Verbeek and Vella (2005), although it is developed using a different approach that is compatible with that of deriving measurement error corrected estimators.

McKenzie (2004) presents an OLS estimator that is similar to our within-group estimators. More specifically, it allows for heterogeneity between cohorts, so (5) can be relaxed to:

$$\bar{y}_{c(t),t} = \bar{\alpha}_c \cdot \bar{y}_{c(t-1),t-1} + \bar{x}'_{c(t),t}\bar{\beta}_c + \lambda_c + \bar{\varepsilon}_{c(t),t} \qquad (20)$$

Under similar assumption, the OLS estimator γ_c^{OLS} for cohort *c* is consistent and has the standard form of:

$$\gamma_c^{OLS} = (U_c'U_c)^{-1}U_c\bar{Y}_c$$

where $\gamma_c = (\alpha_c, \beta_c', \lambda_c)'$; $U_c = (\bar{Y}_{c,-1} \quad \bar{X}_c \quad 1)$. **1** is a vector of ones. $\bar{Y}_{c,-1}$ and $\bar{Y}_c$ both are $(T-1)\times 1$ vector for cohort *c*, whose element is $\bar{y}_{c(t-1),t-1}$ and $\bar{y}_{c(t),t}$ respectively, *t* = 2,…,*T*. $\bar{X}_c$ is a $(\mathrm{T}-1)\times \mathrm{K}$ matrix for cohort *c* with row $\bar{x}'_{c(t),t}$, *t* = 2,…,*T*.

For (20) to be identified, rank condition (8) need to be strengthened to:

$$U_c = (\bar{Y}_{c,-1} \quad \bar{X}_c \quad 1) \text{ has full column rank K + 2.} \qquad (21)$$

This implied that there have to be at least (K + 2) cross sections for the model to be identified. This result is intuitive: when the estimation is done separately for each cohort, the number of cross sections should be at least as large as the number of explanatory variables. In practice, it is unlikely we will be able to reliably estimate the heterogeneous model of (20) with only small number of cross sections, which is probably the most common situation. This will limit the application scope of heterogeneous OLS estimator in empirical work.

5.1.2 Estimator based on individual level data

There are other estimators proposed in the literature, which are designed to use individual data directly without aggregating into cohorts and taking cohort averages. It has been argued that without aggregation, an estimation method based on individual level data can make a more efficient use of the available information (Girma, 2000). In this section, we will discuss two estimators not relying on cohort average and achieving consistency when $n_{ct} \to \infty$. We will show that each has its drawbacks and might not be promising in applied work.

5.1.2.1 Two-Stage Least Square Estimator by Moffitt

Moffitt introduced a two stage least square (2SLS) estimator, where the lagged dependent variable $y_{i(t),t\text{-}1}$ is predicted from the regression on a vector of instrumental variables. First consider an auxiliary regression based on all the sample observations at $t-1$:

$$y_{i(t-1),t-1} = Q'_{i(t-1),t-1}\delta_1 + Z'_{i(t-1)}\delta_2 + \omega_{i(t-1),t-1} \qquad (22)$$

where $Q_{i(t\text{-}1),t\text{-}1}$ is a vector of time-varying variables, and $Z_{i(t\text{-}1)}$ is a vector of time-invariant variables. If $Q_{i(t\text{-}1),t\text{-}1}$ and $Z_{i(t\text{-}1)}$ are strictly exogenous, δ_1 and δ_2 can be consistently estimated using OLS. Using the lagged value $Q_{i(t),t\text{-}1}$ and time-invariant variable $Z_{i(t)}$, the predicted variable $\hat{y}_{i(t),t-1}$ can be obtained from (22). Inserting $\hat{y}_{i(t),t-1}$ in place of $y_{i(t),t\text{-}1}$ in (1) and applying least squares will produce consistent estimates of α, β and λ provided that $\hat{y}_{i(t),t-1}$ is asymptotically uncorrelated with $\varepsilon_{i(t),t}$.

To apply the two stage least square estimator in practice, we need to know the value of the time varying variables in the previous period ($Q_{i(t),t\text{-}1}$), which is a rather strong data requirement. With repeated cross sectional data, the history of the variables is usually unavailable. In practice, $Q_{i(t),t\text{-}1}$ can either be functions of t, which may include the projection of unobserved variables, or variables that can be back-cast with reasonable accuracy. One example of the latter is the number of children in the household, if the ages of children are known.

The 2SLS relies on the prediction of $y_{i(t),t\text{-}1}$ based on functions of t and time-invariant variable Z. If such prediction is done for each individual observation without any

averaging, it is likely that the prediction $\hat{y}_{i(t),t-1}$ will be highly inaccurate. In this sense, the within group estimator of (19) would be preferable as it is based on the cohort average defined by *Z*. Overall, the 2SLS estimator does not seem promising in applied work due to its strong information requirement and likely prediction noise.

5.1.2.2 GMM Estimator of Quasi-differences Model

Girma (2000) proposed a pair-wise quasi-differencing approach for the estimation of dynamic pseudo panel model. Similar to McKenzie (2004), it allows for between cohort heterogeneity, so model (1) can be relaxed to:

$$y_{i(t),t} = \alpha_c y_{i(t),t-1} + x'_{i(t),t}\beta_c + z'_{i(t)}\lambda + \varepsilon_{i(t),t} \qquad t = 2,\ldots,T;\ c = 1,\ldots,C; \tag{23}$$

$y_{i(t),t-1}$ as unobserved variables. Model (23) also implied that the population can be divided into *C* mutually exclusive cohorts and the data exhibit within cohort homogeneity. For any cohort c[18], Girma applied the quasi-difference techniques of

$$y_{i(t),t} - \alpha y_{j(t-1),t-1} \text{ for } \forall[i(t), j(t-1)] \text{ and } \forall t \geq 2$$

and made various assumptions of $y_{i(t),0}$, $x_{i(t),t}$ and $z_{i(t)}$. More specifically, it is assumed that the initial conditions ($y_{i(t),0}$) are random draws with a common (cohort) component in the specification of a conditional mean, and the time-varying variables $x_{i(t),t}$ are AR(1) plus individual specific drift and cohort-time specific error components. This leads to an estimable model of:

$$y_{i(t),t} = \alpha y_{j(t-1),t-1} + \beta' x_{i(t),t} + \eta_{ijt} \tag{24}$$

where η_{ijt} is a composite error term, and two set of linear moment conditions:

$$E\{y_{g(t-s),t-s}\eta_{ijt}\} = 0\ ;\ t = 2,\ldots,T;\ s = 1,\ldots,t-1;\ \forall g(t-1) \neq j(t-1) \tag{25}$$

$$E\{x_{g(t-s),t-s}\eta_{ijt}\} = 0\ ;\ t = 2,\ldots,T;\ s = 0,\ldots,t-1;\ \forall g(t-s) \mid g(t-s) \neq i(t) \text{ or}$$

$$g(t-1) = j(t-1) \tag{26}$$

Conditions (25) and (26) suggest that the any past and present values of the dependent variables and explanatory variables within the same cohort can be used as instruments. Note that in (24) to (26), *i* and *j* index the random sample in different cross sections, so for notational convenience they can be standardized as *i* (since $i(t)$ and $i(t-s)$ represent

[18] For a single cohort, the index on α and β can be dropped.

different individuals sampled so no confusion should arise). To narrow down the potentially infinite number of instruments, (25) and (26) are restricted to become:

$$E\{V_{it}\eta_{it}\} = 0 \tag{27}$$

with $V_{it} = \{y_{g(t-1),t-1}, x'_{g(t-1),t-1}, x'_{g(t),t}\}$, and g is arbitrarily specified as $i - 1$.

The quasi-difference GMM estimator $\gamma_{QDGMM} = (\alpha, \beta')'$ can then be obtained by minimizing:

$$\left[\frac{1}{n(T-1)}\sum_{it} V_{it}\eta_{it}\right]' S_{nT}\left[\frac{1}{n(T-1)}\sum_{it} V_{it}\eta_{it}\right]$$

where S_{nT} is a sequence of weight matrices.

Although the quasi-differencing approach avoids the strong data requirement of Moffitt's 2SLS, it relies on arbitrarily chosen individuals in the same cohort as instruments, which would lead to inaccurate estimation results. More specifically, the lagged value of $y_{i(t),t}$ is approximated by an arbitrarily selected observation $y_{j(t\text{-}1),t\text{-}1}$ in the same cohort and individual observations are used as instruments in the model. As a result, it employs a noisy approximation to the unobserved lagged values as well as noisy instruments (Verbeek and Vella, 2005). Although such noise would cancel out asymptotically, it does not achieve more efficient use of information as Girma claimed. The reason why a different observation *j* would provide information on observation *i* is because they are in the same cohort and this can be more easily captured by the cohort dummy variables, as does the within-group estimators presented above.

5.2 Empirical Results from the Dynamic Car Ownership Model

The pseudo panel dataset constructed from FES is used to estimate the dynamic car ownership model. The full dataset contains 254 observations from 16 cohorts, and after dropping the first observation of each cohort to account for the lagged dependent variable in the dynamic models, the number of observation is reduced to 238. Similar to the static model, the dependent variable is the average number of cars per household in cohort *c* in year *t* ($A_{c,t}$); the explanatory variables include the car ownership number for the same cohort in the previous year ($A_{c,t\text{-}1}$), average household disposable income or its log transformation ($I_{c,t}$), household structure (demographic characteristics)

variables ($S_{c,t}$), average age of household head ($G_{c,t}$), household location variables ($L_{c,t}$) and index of real motoring costs (M_t):

$$A_{c,t} = f(A_{c,t-1}, I_{c,t}, S_{c,t}, G_{c,t}, L_{c,t}, M_t) + \varepsilon_{c,t} \quad (28)$$

Similar to the static model, the household structure variables can take the form of proportions of household as household type 2 to 8 (variable HH2 to HH8, which measure the impact against household type 1, single working person household) [19]; or the average number of children, employed and household size (variable CHILD, WORKER and HHSIZE). The household location variables are either the proportions of households living in four types of area (AREA2 to AREA5, measuring impacts against area type 1, Greater London), or compressed variables for metropolitan areas (variable MET, which combines area type 1 and 2) and the least populated rural area (variable RURAL, same as type 5).

In section 5.1.1.3 of the current chapter, we have shown the Within Group (WG) Estimator is consistent when $n_{ct} \to \infty$. Since we only include cohorts with sufficiently large sample during the construction of the pseudo panel dataset, the WG Estimator is deemed appropriate for the current empirical works. The following sections report the specification search that is mainly based on the WG Estimator.

5.2.1 Assuming linear economic relationship at individual level

The first set of tests assumes the linear economic relationship at individual household level and the data transformations (log, square, etc.) are performed for each household sampled in the Family Expenditure Survey. Each variable in the pseudo panel dataset is also weighted by the square root of the cohort sample size. Such treatments are consistent with the Weighted Least Square Estimator for the static model.

Linear models and semi-log linear models have been estimated. Alternative forms of household structure variables and household location variables (as discussed above) have been tested with each functional form. Standard set of tests applied to each model include the RESET misspecification test, White heteroskedasticity test and Durbin-

[19] See Table 3-2 in Chapter 3 for the definition of household type.

Watson autocorrelation test. High adjusted R square and sensible value of regression coefficients are additional criteria of good model fit.

The models with the best fit include proportions of household types in a cohort as household structure variables, and compressed household location variables. In alternative models, the coefficient for the number of children is consistently negative and significant, possibly due to strong interaction with other household structure variables. The more detailed breakdown of area type does not increase explanatory power of the model and most of the area type variables are not statistically significant. Table 5.1 present the results from the models with best fit, with both linear and semi-log linear forms.

The various specification test statistics are similar for both the linear and semi-log models. Both have high adjusted R square statistic and the RESET test does not reject the hypothesis of no misspecification in both models. Judging by adjusted R^2 alone, one might even conclude that the semi-log model has the better fit. It has turned out that the adjusted R^2 is misleading here.

There are three problems with the semi-log model in Table 5-1. Firstly, the implied adjustment speed is too high to be realistic, which suggests that 93% of adjustment of car ownership to the change of explanatory variables happens in one year. Secondly, the coefficients for purchase price and running costs are significant and of wrong sign. Thirdly, the estimated fixed effects show a wrong trend across cohorts, implying older cohorts with higher car ownership level when other things being equal. Most importantly, all these problems are consistently found in all the eight semi-log models tested (different forms of household structure and location variables; whether or not with log transformation of price and running cost index), and the third problem is also present in the static models discussed in the last chapter. These empirical results seem to suggest that it is not appropriate to assume a linear economic relationship between car ownership and various explanatory variables at individual household level.

Table 5-1 Models with best fit (comparison of linear and semi-log functional form)

	Linear		Semi-Log	
	Coeff	**t-ratio**	**Coeff**	**t-ratio**
CAR [-1]	0.1482	4.00	0.0740	2.17
INC	0.0004	3.14		
ALINC			0.3963	8.61
HH2	-0.2069	-1.09	0.1465	0.84
HH3	-0.4633	-1.77	0.0130	0.05
HH4	0.4687	3.09	0.5694	4.27
HH5	0.4964	2.69	0.5342	3.27
HH6	0.5939	3.89	0.4256	3.11
HH7	0.9901	5.16	0.9481	5.60
HH8	0.7925	4.36	0.6357	3.99
MET	-0.1497	-1.43	-0.1981	-2.15
RURAL	0.2134	1.93	0.1116	1.15
PRICE	-0.0006	-0.76		
RUNCST	-0.0016	-3.07		
LNPRICE			0.1179	1.64
LNRUNCST			0.1598	2.64
AGE	0.0351	8.26	0.0160	3.55
AGSQ	-0.0002	-4.83	-0.0002	-6.15
C1	-1.1933	-5.16	-2.8161	-4.96
C2	-1.1316	-5.07	-2.8459	-5.00
C3	-1.1046	-5.03	-2.9143	-5.09
C4	-1.0389	-4.82	-2.9638	-5.13
C5	-0.9865	-4.65	-3.0131	-5.17
C6	-0.9132	-4.38	-3.0483	-5.17
C7	-0.8151	-3.99	-3.0454	-5.12
C8	-0.7303	-3.65	-3.0467	-5.07
C9	-0.6157	-3.16	-3.0298	-4.99
C10	-0.5340	-2.80	-3.0465	-4.96
C11	-0.4662	-2.51	-3.0800	-4.96
C12	-0.3862	-2.12	-3.1122	-4.94
C13	-0.2989	-1.68	-3.1360	-4.92
C14	-0.2100	-1.20	-3.1555	-4.88
C15	-0.1693	-0.97	-3.2183	-4.92
C16	-0.1686	-0.98	-3.2989	-5.01
Adjusted R^2	0.993		0.995	
SSE	114.27		88.43	
Log Likelihood	-250.4		-219.89	
DW stat	2.04		2.11	
F-stat of White Test	5.72		3.22	
t-stat of RESET Test	-0.54		0.11	

Note: $y = A_{ct}$ in both models

On the other hand, the results from the linear model are more sensible[20]. Based on the linear model, the short run income elasticity is 0.14 and running cost elasticity is -0.17 for mid-income household; the corresponding long run elasticity is 0.17 for income and -0.2 for running cost. Compared to the income elasticity of 0.43 (also for mid-income household) derived from the static linear model, both the short run and long run elasticities are much lower[21]. In general, the elasticities derived from the dynamic model are lower than the ones estimated by Dargay and Vythoulkas (1999), especially the long run elasticities. This is largely due to the much smaller coefficient for the lagged dependant variable (LDV) in the current work. The explanation and implication of small LDV coefficient will be further explored in the next sub-section.

The coefficients for the variables of 'Age' and 'Age Square' are positive and negative respectively, indicating a peak in the life cycle of car ownership. Regarding the coefficients of household structure and location variables, they are not very different from those estimated by the static model, so no further discussion is presented here.

5.2.2 Assuming Linear Economic Relationship at Cohort Level

As discussed in the previous chapter, there are theoretical and practical problems in assuming a linear economic relationship of car ownership and explanatory variables for each individual household. Theoretically, it is unrealistic to assume car ownership growth is linear at household level and this assumption is particularly problematic in the dynamic setting. This is because for individual household, the increase of car ownership level is not linear, and the relationship between the past and current car ownership can not be simply measured by a constant parameter. The empirical results reported earlier do not support the assumption of a linear relation at household level. Furthermore, we can not directly forecast future values of the explanatory variables if the log transformation is done before averaging into cohorts so such a model becomes useless for forecasting purpose.

---

[20] As it does not involve any log transformation, linear model can be used to describe economic relationship at either individual household level or cohort level. No distinction can be made between these interpretations and the linear model results are discussed here purely for convenience of presentation.

[21] For the dynamic model, the coefficient for purchasing price variable is not significant so no reliable elasticity can be derived; for static model, the coefficient for running cost variable is not significant. As a result, comparison is only done for income elasticity.

Alternatively, we can view the economic relationship between car ownership and the explanatory variables as linear at cohort level. In this case, the transformation (log, square, etc) of variables should be done on cohort average. As discussed in the previous chapter, there are some advantages of weighting all the variables by the square root of the sample size of each cohort, as the observations in the synthetic panel that are more accurately measured will be assigned a greater weight. As a result, similar weighting is done for the dynamic model.

5.2.2.1 Specification Search

The specification search is similar to that described in the previous section, although we also investigate models with double log functional form here. The coefficient for the CHILD variable (average number of children) is consistently negative and significant, which is opposite to expectation. The coefficient for the WORKER variable (average number of person in work) is positive but not significant for all linear and semi-log linear models; for all double log models, it is negative and significant. However, the double log models with CHILD, WORKER and HHSIZE as household structure variables could be mis-specified according to the RESET test. On the other hand, using the split of household types as explanatory variables produces more satisfactory results. Regarding the household location variables, the more detailed breakdown of location type does not offer additional explanatory power to the model with only the AREA5 variable (proportion living in the least populated rural area) being significant. Consequently, the compressed location variables are chosen in the preferred model.

For the models with the preferred form of household type and location variables, the difference between those with different functional forms (linear, semi log and double log) is more subtle. All models have high adjusted R square, and none of the RESET test rejects the hypothesis of no misspecification. The Durbin-Watson statistic is very close to 2 for all models, which does not indicate the presence of autocorrelation. However, the White test rejects the hypothesis of homoskedasticity for all models, which will be addressed later.

Similar to the static model, the sign and magnitude of the regression coefficients are used as additional criteria to determine the model of best fit. The household structure

and location variables are the same for all three models, and their coefficients are very similar for the linear and semi-log models (and comparable for the double log model). None of the coefficients of the purchase price elasticity are significant at 10% level, so no reliable conclusion can be made about the price elasticity. The income and running cost elasticities, as implied by models of different functional forms, are different. Table 5.2 and 5.3 compare the short run and long run elasticities implied by the three models.

Table 5-2 Short Run and Long Run Income Elasticity

Income / Car	Short Run			Long Run		
	Linear	Semi-Log	Dbl-Log	Linear	Semi-Log	Dbl-Log
Low	0.173	0.394	0.225	0.203	0.460	0.316
Middle	0.141	0.181	0.225	0.165	0.211	0.316
High	0.145	0.133	0.225	0.171	0.155	0.316

Table 5-3 Short Run and Long Run Running Cost Elasticity

Car	Short Run			Long Run		
	Linear	Semi-Log	Dbl-Log	Linear	Semi-Log	Dbl-Log
Low	-0.371	-0.309	*-0.067*	-0.436	-0.360	*-0.095*
Middle	-0.170	-0.141	*-0.067*	-0.200	-0.165	*-0.095*
High	-0.125	-0.104	*-0.067*	-0.147	-0.121	*-0.095*

Note: 1. Low, middle and high real disposable income are 172, 306 and 430 pound per week respectively; low, middle and high car ownership level are 0.42, 0.92 and 1.25 cars per household respectively.

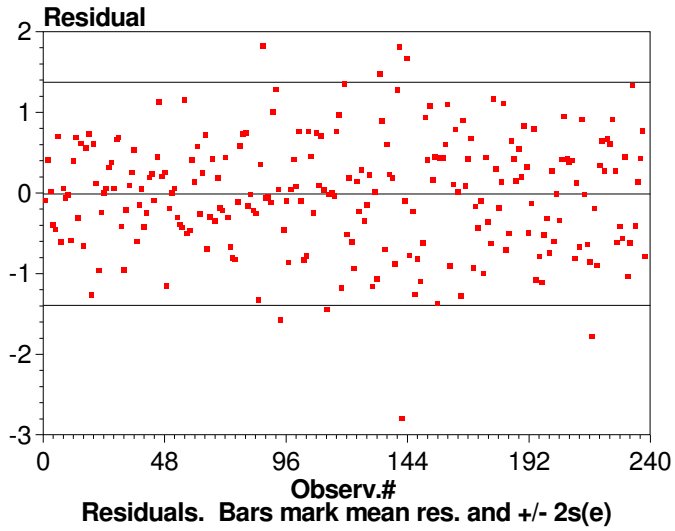
2. The running cost elasticity is based on a constant cost index of 100 (1995 level);

3. The coefficient of running cost variable is not significant for the double log model, and hence the corresponding elasticity is in italic.

In the double log model, the income and running cost elasticities are constant for families with variable income/car ownership level. This is a characteristic of the double log model and could be problematic for car ownership models, where the income elasticity is known to decline with income and "saturation" is observed in mature car markets such as the UK. Consequently, the double log model is regarded as inappropriate form. The income elasticities implied by the linear model are low and similar cross household with various income and car ownership level; meanwhile, the running cost elasticity is much higher than the income elasticity for low car ownership households and declines rapidly with car ownership level. Regarding the semi-log model, the income elasticity declines rapidly with car ownership level, and the running cost elasticities are always lower than the income elasticities by a consistent proportion. Overall, the income and price elasticity implied by the semi-log model seem more sensible so the semi-log model is selected as the preferred model.

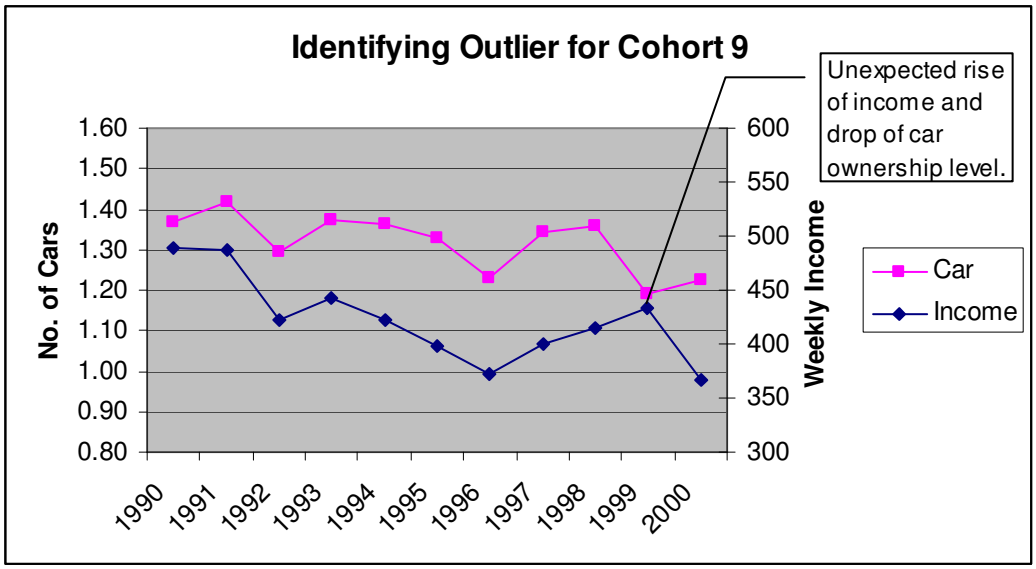
As mentioned above, White test detects presence of heteroskedasticity in all models with different functional forms. Further examination of the residual identifies that one observation has particularly high prediction error. Figure 5.1 shows the residual plot of the semi-log model with the preferred household structure and location variables.

Figure 5-1 Residual Plot of Semi-Log Model with outlier



Note: X-axis is ordered by year for each cohort

Figure 5-2 Identifying outlier for cohort 9 in year 1999



The outlier is the observation for cohort 9 (household heads born between 1941 and 1945) in 1999. The real weekly disposable income increased from £415 in the previous year to £433 and then suddenly dropped to £367 in the following year; on the other

hand, the car ownership level dropped from 1.36 cars per household in 1998 to 1.19 cars and then increased to 1.23 cars. Figure 5.2 illustrates the unexpected change of income and car ownership level for cohort 9 in 1999.

In the dynamic model, the outlier can not be simply excluded, since that will upset the dynamic relationship. As a result, a new dummy variable "OUT", which takes the value of 1 for the outlier and 0 for other observations, was added to the model. The coefficient of the "OUT" variable measures the prediction error for this observation. The likelihood ratio test produces a Chi Square statistic of 21.7, which strongly suggests the increased level of fit with the additional variable. When the model is re-estimated, the coefficient for the purchase price variable is wrong signed and not significant. As a result, the purchase price variable is dropped from the model.

5.2.2.2 Results of the Preferred Models

The specification search identifies the most suitable functional form and explanatory variables. It is also discovered that there is a strong linear trend in fixed cohort effects, similar to the static model. As a result, a restricted model is estimated, where the cohort dummy variables are replaced by a variable of cohort ID. Table 5.4 reports the results of the unrestricted and restricted fixed effect model with the semi-log functional form.

The likelihood ratio test leads to a chi square statistic of 54.3, which strongly favour the unrestricted fixed effect model. The RESET test rejects the hypothesis of no misspecification for the restricted FE model, and the coefficient of the log purchase price variable is significant and with wrong sign. On the other hand, the RESET test does not indicate misspecification for the unrestricted FE model. Its residual plot (Figure 5.3) does not show any apparent misspecification.

The coefficient of the lagged dependent variable (LDV) in the unrestrictive fixed effect model is 0.19. Based on the structure of exponentially distributed lag, the implied adjustment speed is 81% in one year and the full adjustment takes mere four years (99.9%). Why is the adjustment speed here is much higher than that reported in Dargay and Vythoulkas (1999)? To understand the impacts of model specification on the coefficient of LDV, we report the results from 4 comparable models in Table 5-5.

Table 5-4 Unrestricted and restricted fixed effect model (semi log form)

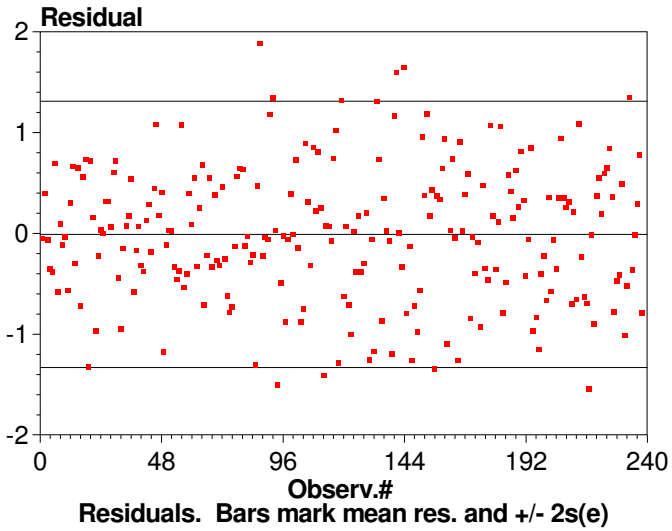
	Unrestricted FE		Restricted FE	
	Coeff	**t-ratio**	**Coeff**	**t-ratio**
Constant			-2.7679	-4.08
CAR [-1]	0.1889	5.29	0.2903	7.79
LNINC	0.1894	4.20	0.2336	4.88
HH2	-0.1090	-0.60	-0.4286	-2.43
HH3	-0.3393	-1.35	-0.6618	-3.19
HH4	0.4702	3.31	0.2035	1.40
HH5	0.4802	2.78	0.1222	0.70
HH6	0.5201	3.65	0.2497	1.72
HH7	0.9149	5.14	0.5430	3.00
HH8	0.7286	4.31	0.4493	2.66
MET	-0.1612	-1.61	-0.3184	-3.14
RURAL	0.1942	1.84	0.2637	2.36
LNPRICE			0.1788	2.27
LNRUNCST	-0.0958	-2.08	-0.0235	-0.43
AGE	0.0306	7.29	0.0173	5.02
AGSQ	-0.0002	-4.53	-0.00005	-1.41
COHORT			0.0694	8.22
C1	-1.7403	-4.91		
C2	-1.6816	-4.79		
C3	-1.6537	-4.67		
C4	-1.5985	-4.47		
C5	-1.5547	-4.30		
C6	-1.4920	-4.08		
C7	-1.4019	-3.80		
C8	-1.3214	-3.56		
C9	-1.2043	-3.22		
C10	-1.1352	-3.00		
C11	-1.0751	-2.81		
C12	-1.0067	-2.60		
C13	-0.9314	-2.38		
C14	-0.8541	-2.16		
C15	-0.8214	-2.04		
C16	-0.8104	-2.00		
OUT	-0.1554	-4.44	-0.1559	-4.19
Adjusted R^2	0.994		0.993	
SSE	103.78		130.44	
Log Likelihood	-238.94		-266.15	
DW stat	2.05		1.88	
t-Stat of RESET test	-0.23		2.03	

Note: $y = A_{ct}$ in both models

Table 5-5 Model Specification and LDV Coefficient

	Fixed Effect	Restrictive FE	Rst FE, HH_Var * 3	Rst FE, HH_Var * 3, no Age, AgSq
Adjusted R^2	0.994	0.993	0.992	0.991
LDV Coeff	0.19	0.29	0.34	0.44

Figure 5-3 Residual Plot of the unrestricted fixed effect model



Note: X-axis is ordered by year for each cohort

Comparing the 4 models in Table 5-5, one can see that while the goodness of fit gets worse, the coefficient for the LDV increases. The first two models are those reported in Table 5-4. The larger LDV coefficient in the restrictive model might be a reflection of uncontrolled cohort heterogeneity. This interpretation is an analogy to the phenomenon of 'spurious state dependence' in the discrete choice model to be discussed in Chapter 7. The difference between the second and the third model lies in the form of household characteristic variables. When the average number of children, people in work and household size are used as explanatory variables (rather than split of 8 household types), the LDV coefficient goes up again. In the fourth model, when the variables of 'Age' and "AgSq' are dropped and the resulting specification is similar to Dargay and Vythoulkas (1999), the LDV coefficient becomes close to their figures. Actually, their original estimates seem to have substantial upward bias due to omission of relevant variables or other mis-specifications, as the more general models in Dargay (2001) and Dargay (2002), which consider more flexible functional form, segmentation of explanatory variables and unrestrictive cohort effects, produce much smaller coefficient for LDV.

Returning to the unrestricted fixed effect model in Table 5-4, the coefficients for the log income and log running costs variables are significant and of expected sign. The implied income elasticity and running cost elasticity are comparable to those from the

corresponding static model (the coefficient of purchasing price variable is not significant in both models). Table 5.6 compares the short run and long run elasticities from the dynamic model with elasticities from the static model.

Table 5-6 Comparison of elasticities (dynamic and static unrestricted FE models)

Car	Income Elasticity			Running Cost Elasticity		
	Dynamic SR	Dynamic LR	Static	Dynamic SR	Dynamic LR	Static
Low	0.461	0.569	0.560	-0.228	-0.281	-0.339
Middle	0.211	0.261	0.256	-0.104	-0.129	-0.155
High	0.155	0.191	0.188	-0.077	-0.095	-0.114

The coefficients of the household structure variables, location variables, and second polynomial of the age of household head variables are all comparable to those of the static model. All the coefficients for the proportion of households with two or more adults are positive and significant. According to the model, for example, if the proportion of households with three or more adults but no children (Household type 7) increases by 1%[22], the average number of cars per household would increase by 0.0091 for that cohort; similarly, if the proportion of households with three or more adults plus children (Household type 8) increases by 1%, the average number of car would increase by 0.0073. The parameter for the proportion of living in metropolitan area is negative (although not significant), and that for the proportion of living in the least populated rural area is positive. The value of the coefficient suggests that a 1% increase of households living in the least populated rural area would increase the average car number by 0.0019 per household in the cohort.

5.2.2.3 Alternative Model Specification and Estimation

Similar to the static models, we also test a set of random effect models. While the Hausman's Test indicates preference to the fixed effect model in only about half of the models estimated, the RESET test rejects the hypothesis of no misspecification in almost all models. This is likely to be caused by the correlation between the explanatory variables and the cohort effects, which would violate the orthogonality assumption of the random effect model.

[22] This implies the proportion of single working household decreases by 1%.

We also investigate the hypothesis of homogeneity by estimating cohort specific models. Due to insufficient number of observations, the youngest and two oldest cohorts have to be dropped. To save degrees of freedom, the household structure variables are average number of children and household size, and compressed location variables are used. Then, the models are estimated for each of the 13 cohorts separately. The results are unsatisfactory, as the majority of the coefficients are not significant even at 10% level. It is unlikely that we are able to make reliable inference on elasticity or to make forecasts based on these models.

Finally, we investigate the robustness of the estimation using the parametric bootstrap technique, which is implemented in four steps. First, the unrestricted fixed effect model in semi-log form (as reported in Table 5.4) was estimated, and the predicted value $\hat{A}_{c,t}$ was saved. Second, a series of random number *W* was generated with zero mean and standard deviation of 0.65, which are the approximate moments of the residual from the model estimated in step one. Third, the model is re-estimated with a simulated dependent variable $A^S_{c,t} = \hat{A}_{c,t} + w$, where *w* are random draws from *W*. It should be noted that for the first observation in each cohort, the simulated variable is $A^S_{c,t0} = A_{c,t0} + w$, where $A_{c,to}$ is the "observed" value in the pseudo panel dataset rather than the predicted value[23]. Finally, step three is repeated 1000 times and the coefficients estimated from each run are saved for further analysis.

In general, the coefficients estimated using "observed" dependent variable are consistent with those based on simulation. Figure 5.4 plots the distribution of the simulated coefficients for the log income variable, compared with the value of the coefficient estimated based on "real" data. The latter is close to the centre of the bell shape distribution, which shows that the initial estimate is broadly unbiased.

Figure 5.5 shows the distribution of the simulated coefficient for the log running costs variable as well as the point estimate based on observed data. Similar to above, the

[23] $A^S_{c,t0}$ is only used as lagged dependent variable (regressor).

latter is very close to the centre of the simulated distribution. This also confirms the unbiasedness of the point estimate coefficient for the log running cost variable.

Figure 5-4 Log Income Variable: distribution of the simulated coefficients and point estimate based on real data

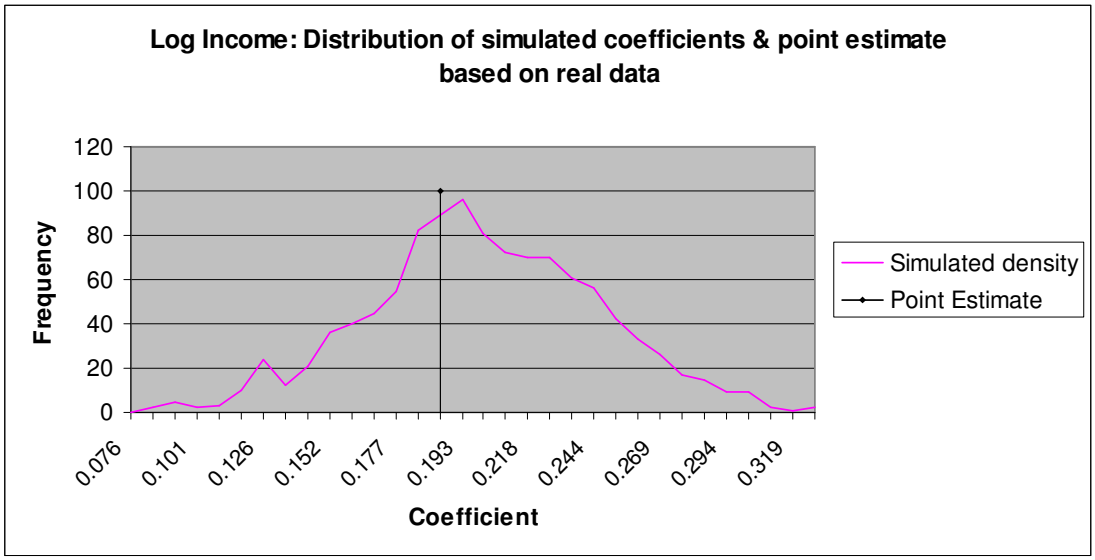


Figure 5-5 Log Running Cost Variable: distribution of the simulated coefficients and point estimate based on real data

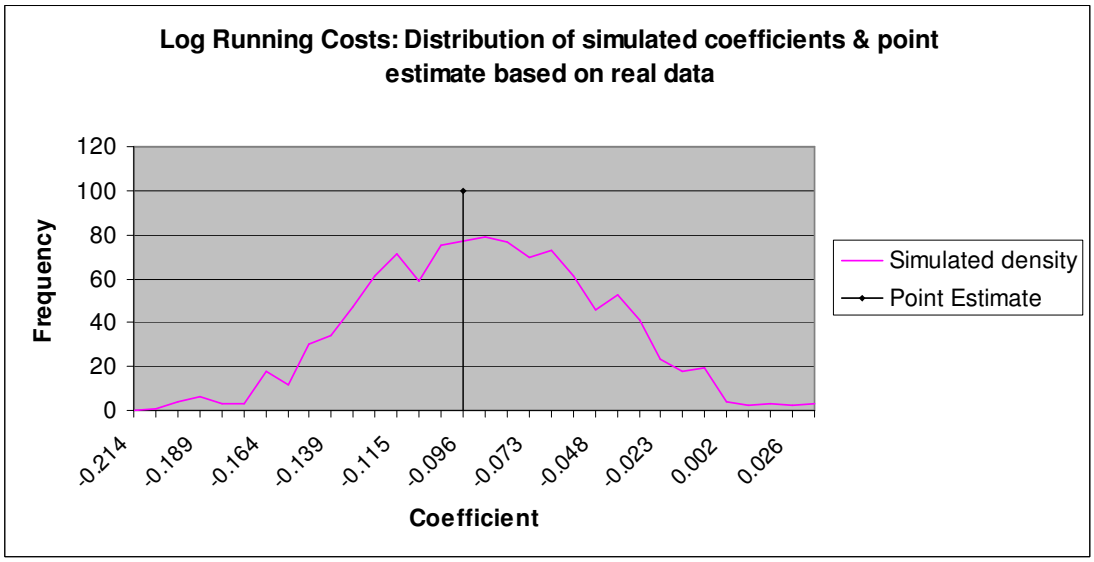
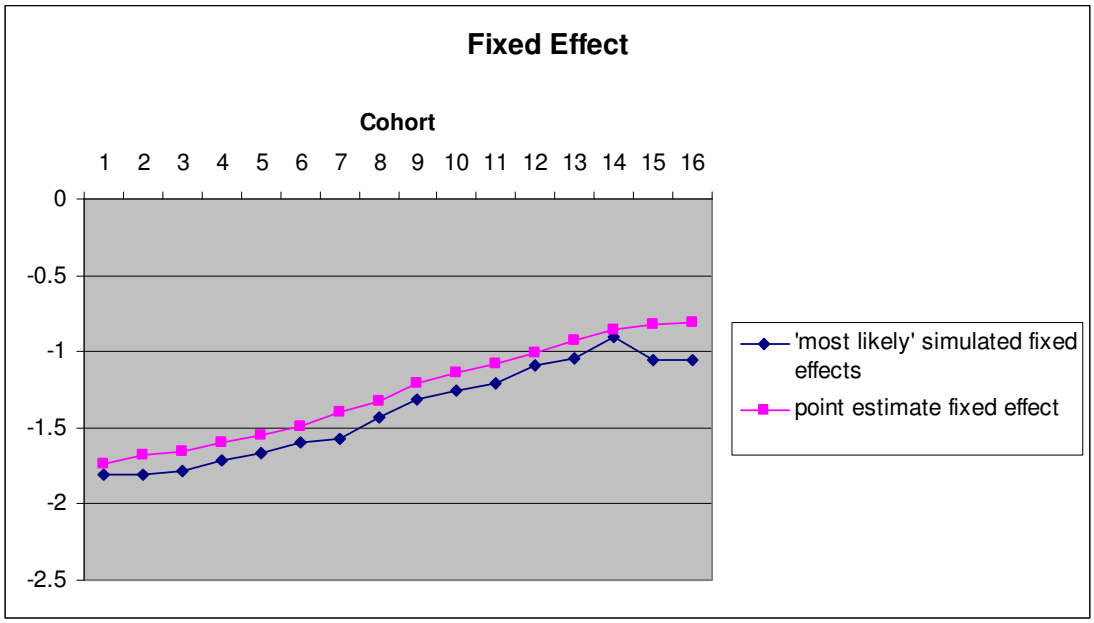


Figure 5.6 compares the "most likely" fixed effects from simulation and the point estimate based on observed data. To obtain the "most likely" simulated coefficient, we divide the 1000 parameter values obtained from simulation into 31 ranges, and determine the "most likely" range, which is the one with the most occurrences. The mid-point of the "most likely" range becomes the representative simulated fixed effect. Figure 5.6 shows the point estimates based on real data are quite close to the "most

likely" simulated values expect for the youngest cohort. As discussed before (and in the literature, e.g. Dargay and Vythoulkas, 1999), the linear trend in the fixed effect become less clear for the youngest cohort, and this effect seems to be amplified with the simulated data.

Figure 5-6 "Most Likely" fixed effects from simulation and point estimate from real data



5.3 Conclusion

In this chapter, we first investigate the theoretical aspects of the dynamic pseudo panel. Two types of consistent estimators have been discussed: one based on cohort average and the other based on individual survey data. Regarding the former, we review the Error Corrected Within-Group Estimator, consistent when $T \to \infty$, and the Error Corrected GMM Estimator, consistent when the number of cohort is large ($C \to \infty$), both of which were proposed by Collado (1997). We also present a Within-Group Estimator, which is computationally attractive and consistent when the number of sample observation is large for each cohort unit ($n_{ct} \to \infty$). Certain rank conditions have to be satisfied for identification, which requires the cohort means of the dependent and independent variables should not exhibit perfect collinearity and vary over time. It is also required that there are at least three cross sections for the model to be identified.

Regarding the estimators based on data at individual level, we review the Two-Stage Least Square Estimator by Moffitt (1993) and the GMM Estimator of Quasi-

Differences model by Girma (2004). Both approaches involve noisy approximation of the lagged dependent variable. Although such noise would cancel out asymptotically, its impacts on practical application might be severe. In another word, their usefulness in empirical work would be limited.

The second part of this chapter reports the empirical work on dynamic car ownership model. As the pseudo panel is a synthetic panel constructed from individual survey data, we can assume a linear economic relationship between the car ownership level and various explanatory variables either at individual household level or at cohort level. Consequently, separate sets of the specification search have been carried out to determine the model with best fit under each assumption.

Assuming a linear economic relationship between the dependent variable and the regressors for each household requires the transformation (log, square, etc.) of individual survey data before obtaining cohort average. However, the empirical results from the semi-log models show various problems including a very low coefficient for the lagged dependent variable, coefficients with wrong sign for motoring cost variables and wrong trend in fixed effects across cohorts. These problems reinforce the idea that it is not appropriate to view the relationship between car ownership and explanatory variables as linear for each individual household, especially when the regressors include a lagged dependant variable. This is because for individual household, the relationship between past and current car ownership level is unlikely to remain constant over time.

Alternatively, we assume that the economic relationship is linear at cohort level. A number of models have been estimated, with different explanatory variables, functional forms and representation of cohort effects. Systematic specification search has been carried out, and an un-restrictive fixed effect model in semi-log form is found to have the best goodness of fit. The implied long run income elasticity is 0.57, 0.26 and 0.19 for households with low, middle and high car ownership level; the implied long run running cost elasticity is -0.28, -0.13 and -0.10 respectively, which all appear to be sensible. To check the robustness of the estimation, the preferred model is re-estimated using parametric bootstrap technique, and the distributions of the simulated coefficients confirm the unbiasedness of the point estimates obtained using the pseudo panel data.

Chapter 4 and 5 thoroughly investigate the static and dynamic car ownership model with linear (or generalised linear) form. The models with the best fit will be used for forecast at a later stage. The next two chapters will investigate the car ownership models with non-linear form.

Chapter 6 Random Utility Model of Pseudo Panel

In demand forecasting, the presence of non-linearity could have significant effects on the accuracy of the forecast results. In the previous chapters, the misspecification tests do not appear to reject the hypothesis of linearity for many of the estimated models. However, this does not preclude further investigation on nonlinear pseudo panel models. Nonlinear pseudo panel models can take the functional form of Logit, Probit or other forms of discrete choice model, while the underlying data are average characteristics of the cohort sample. As a result, it becomes possible to investigate the impact of dynamics and saturation within a single modelling framework. Such models can be regarded as a "third way" of analysing repeated cross section data, presenting advantages over the linear pseudo panel models and the cross sectional discrete choice models.

This chapter is organised as follows: section one introduces the nonlinear pseudo panel model and discusses its advantage and disadvantage by comparing it to the linear pseudo panel model and cross sectional model. Section two develops a random utility model of pseudo panel and as an application, a hierarchical car ownership model. Section three discusses the consistent estimation of the discrete choice pseudo panel model including models with fixed effect and random effect, drawing from a growing econometric literature on nonlinear (genuine) panel model. Section four reports the empirical results for the car ownership model, although limited to static models of various forms. Dynamic models and models with saturation will be discussed in the following chapter.

6.1 Pros and Cons of Nonlinear Pseudo Panel Models

As the pseudo panel model is based on the cohort average data, it normally has linear form. For models of durable goods or other discrete choice, linear pseudo panel would be regarded as an approximation of economic relationship at (cohort) aggregate level. While such an approximation appears to be sufficient in most empirical studies, it could be beneficial to extend the pseudo panel technique to a discrete choice model. In this section, we will discuss the pros and cons of nonlinear (more specifically, discrete choice) pseudo panel model and argue for its potential as an effective "third way" in

modelling and forecasting using repeated cross sectional data. First, comparison will be made between nonlinear and linear pseudo panel models, evaluating the advantage and disadvantage of each approach. Then similar comparison is made between nonlinear pseudo panel and cross sectional models.

6.1.1 Nonlinear and Linear Pseudo Panel Models

In the literature of durable goods, saturation is an important concept. It is a limit on the choices faced by decision maker, which may be reached but not exceeded. In the linear pseudo panel model, saturation can only be implicitly handled by choosing the appropriate functional form, e.g. in semi-log models the elasticity declines with the rise of income. On the other hand, the impact of saturation can be explicitly considered and estimated, and its statistical significance can also be examined in nonlinear pseudo panel models. By specifying car ownership models with an S-shape functional form and a saturation level, forecasts of vehicle ownership will be curtailed as saturation is approached. Although probably not being significant in developing countries, this feature would be highly significant to forecasts in more mature passenger car markets such as Great Britain (Whelan et al, 2000).

In chapter 4, we discuss the problem of interpreting the Weighted Least Square Estimator as the Instrumental Variable estimator of individual household data. In that case, the number of cars owned or used by the household is treated as a continuous variable, which is inconsistent with consumer's utility maximization behaviour on durable goods. Alternatively, the linear pseudo panel model can be interpreted as an aggregate model, with the cohort average as the unit of observation. In that case, this model becomes detached from the microeconomic theory of utility maximization.

On the other hand, nonlinear pseudo panel models can be specified within the framework of the Random Utility model, thus ensuring their consistency with the economic theory. As will be shown in the following section, the utility function of car ownership can be specified as a deterministic term based on the mean sample characteristics of households in each cohort plus various random components. In this way, the model would be based on economic theory rather than aggregate empirical functions.

The nonlinear pseudo panel model also has its shortcomings. Firstly, the fixed effect models suffer the "incidental parameter" problem and can not be consistently estimated using linear panel model's demeaning technique. This is a problem suffered by nonlinear panel model in general and will be discussed in more details later in this chapter. Secondly, the incidental parameter problem is complicated by the measurement error problem, making it difficult to establish the consistency conditions under various asymptotics. Unlike the linear models, where the theory is relatively well developed (see Chapter 4 and 5 for detailed reviews), there are no theoretical studies on nonlinear pseudo panel models to the author's knowledge. Thirdly, for empirical work, only some basic models can be estimated using commercial econometric packages, while more advanced models such as random parameter logit models (also called mixed logit models) are beyond the reach of readily available software. The current study is the first attempt to address some of these issues, although further research into various theoretical and empirical aspects of nonlinear pseudo panel models is likely to produce more fruitful results.

6.1.2 Pseudo Panel and Cross Sectional Models

After comparing the nonlinear pseudo panel model with its linear counterpart, we make a similar comparison to the cross sectional discrete choice models. As modelling "third way", nonlinear pseudo panel model also has two advantages over cross sectional models. The first is the inclusion of dynamics in modelling and the second is effective tackling of the aggregation bias problem.

The static cross sectional models rely on the assumption of equilibrium, which in practice is the exception rather than norm. The disequilibrium status might be revealed by the instability of cross sections, i.e. different parameter estimates are obtained using cross sectional data in different years. If we believe that each cross sectional sample is representative of the population and long run equilibrium exists, then the cross sectional instability can be explained by the divergence of each cross section from such equilibrium. As the divergence depends on the determining factors in the current and previous periods, the degree of disequilibrium will vary between years, so will the parameter estimates (Dargay and Vythoulkas, 1999).

When the cross sectional data are not in equilibrium, it can no longer be assumed that such data capture the long run relationship[24], and the model based on them will produce biased estimates of long run parameters. When these biased parameters are used for policy analysis, it can lead to wrong conclusions; when they are used in forecasts, it can lead to biased results. However, all these problems could be tackled in the pseudo panel setting, where the dynamic effects can be explicitly quantified and analyzed. In a dynamic model, it is possible to examine the significance of the lagged effects, the speed of adjustment, the extent of asymmetry[25], etc. Long-run as well as short run elasticity, which has significant policy and practical importance, can be obtained from dynamic models. The unbiased parameters derived from the dynamic models could in theory improve the performance of the forecasting models as well.

Another potential advantage of nonlinear pseudo panel model relates to the choice of aggregate and disaggregate model. In many practical applications, including car ownership forecasts in the current study, the subject of interest is the aggregate statistics. Traditionally, there are two approaches to obtain aggregate measures such as market shares from data at individual level, i.e. aggregating individual data either before or after model estimation. Various classical aggregate models belong to the first approach, which are subject to various criticisms including inefficiency in the use of data, not accounting for full data variability and the risk of statistical distortion such as ecological fallacy (Ortuzar and Willumsen, 2001). Although the second approach addresses most of these criticisms, the difficult question of how to perform aggregation based on micro relations remains.

The simplest method, naïve aggregation of the discrete choice model, uses average characteristics of the individual (household) to forecast the aggregate choice probability or market share. It is well known that such an approach gives biased results, and consequently, a number of alternative approaches have been proposed (Ben-Akiva and Lerman, 1985; Ortruza and Williamsum, 2001; Whelan, 2003). Among them, the most robust approach is sample enumeration, where the choice probability of each individual is averaged over all observations within the sample. If the sample is

[24] Except for some very special cases, e.g. homogenous cointegration between variables at the individual level (Madsen, 2005).
[25] A notable example on car ownership is Dargay (2001).

representative of the population, this approach will give unbiased estimate of the aggregate estimate. However, it encounters difficulties in the long term when the distribution of attributes in the population will be different from the base year sample. In this case, the base year sample needs to be adjusted so that it can be considered representative. Daly and Gunn (1985) proposed a method called prototypical sample enumeration, which involves creating an artificial sample with the same aggregate characteristics of those forecasted by planners (e.g. age and sex distribution of the populations). Another method of aggregation is the classification approach, i.e. classifying the sample into homogeneous groups and using group average characteristics as input to the discrete choice model. The accuracy of this method depends on the number of classes and their selection criteria, and good methods of defining the classes were suggested by McFadden and Reid (1975). Nevertheless, this method still involves using average characteristics as input to the disaggregate model and some degree of aggregation is inevitable.

The nonlinear pseudo panel model is the "third way" in obtaining aggregate statistics from data at individual level. Individual data are aggregated into homogenous groups (cohorts), and the models are estimated using average characteristics of the cohort sample. As a result, the empirical model describes the economic relationship between the observed share of choices and explanatory variables at the cohort level. The probability of (cohort) decision-makers making a certain choice, when estimated based on such a model, would give an unbiased estimate of the market share for that choice[26]. Moreover, the explanatory variables are cohort average characteristics that could be directly derived from published planning statistics, thus avoiding the need for more complicated procedures such as prototypical sample enumeration at the forecasting stage. This feature would make nonlinear pseudo panel particularly attractive for long term forecasting based on cross sectional data. For longer term forecasts, Daly and Ortuzar (1990) conclude that the use of more aggregate data tend to have more favorable cost/accuracy trade-off; the discussion here suggests that using pseudo panel data has the potential of actually improving the forecast accuracy.

[26] For multinomial logit model whose utility function includes a constant term, this result directly follows the first order condition of the log likelihood function.

Nevertheless, the aggregation of cross sectional data into cohorts thus reduces the variability of the data, a similar criticism suffered by the aggregate models. For car ownership model, it might cause difficulty in the estimation of saturation levels (although it is not the case for the current study). Furthermore, the information on individual decision makers will be lost after aggregation, and only the average characteristics of cohort sample remain observable. The discrete choice pseudo panel model would have a composite error term, which makes the model parameters estimated on pseudo panel data not directly comparable to those estimated on individual data due to the different scale. This issue will become apparent in the discussion of the random utility model in the next section. Generally speaking, one should probably be more cautious in using nonlinear pseudo panel models as analytical tools, as whether the various disadvantages are outweighed by the inclusion of dynamics remains to be seen.

We summarise the above discussions in Table 6.1, highlighting the advantage and disadvantage of nonlinear pseudo panel models compared with the linear models and cross sectional models.

Table 6-1 Advantage and Disadvantage of nonlinear pseudo panel model

	Vs. Linear Pseudo Panel Model	*Vs.* Cross Sectional Model
Advantage	• Explicitly modelling and estimating saturation level; • Consistent with theory of utility maximization;	• Consideration of dynamics in modelling; • Effective tackling of aggregation bias problem;
Disadvantage	• Bias in the Fixed Effect Estimator; • Lack of ready-made software for advanced models;	• Reduction of data variability; • Loss of information on individual decision makers;

6.2 A Random Utility Model of Car Ownership

In the previous section, we present the nonlinear pseudo panel as a "third way" between its linear counterpart and cross sectional discrete choice model. However, to the best of our knowledge, there is no previous application of this method. One possible explanation is the perceived contradiction between the disaggregate nature of the discrete choice model and the fact that pseudo panel is aggregated from cohort

sample. In this case, it is tempting to conclude that pseudo panel is an aggregate model so the use of discrete choice method is not appropriate. Such belief seems misguided.

In this section a simple Random Utility Model will be presented. We start from a general random utility model for pseudo panel; then proceed to the car ownership model in terms of model structure and specification of utility function; and finally we discuss the estimation of such model. It should be noted that the model presented here is in a general form without explicit consideration of dynamics, which will be discussed in the next chapter.

6.2.1 Random Utility Model of Pseudo Panel

The concept of utility was initially introduced by the neoclassical economic theory of consumers. The consumer is assumed to have preferences on all the possible alternatives *a, b…* (consumption bundles in the neoclassical literature) in the choice (consumption) set ***A***. We note $a \succeq b$ if *a* is at least as good as *b*. Assuming that preferences are complete, reflexive, transitive, continuous and strongly monotonic, then there exists a continuous utility function *U*, which represents those preferences (detailed discussion of these assumptions can be found in Varian, 1992):

$$a \succeq b \Leftrightarrow U(a) \geq U(b) \qquad \forall a, b \in A \qquad (1)$$

Because the choice set ***A*** is finite, there must exist an alternative a^*, which is most preferred to (or as good as) the rest of them:

$$a^* \succeq a \Leftrightarrow U(a^*) \geq U(a) \qquad \forall a \in A \qquad (2)$$

So the most preferred alternative a^* is identified by:

$$a^* = \arg\max_{a \in C} U(a) \qquad (3)$$

In the neoclassical setting, the preference maximization problem is defined subject to a budget constraint and the solution exists when the utility function is continuous and the constraint set is closed and bounded. However, the strong assumptions required by the neoclassical economic theory severely limit its practical application. The complexity of human behaviour suggests that a choice model should explicitly capture some level of

uncertainty. The Random Utility Model inherits the deterministic decision rules from the neoclassical economic theory, while capturing uncertainty by the random components of the utilities (Bierlaire, 1998).

The random utility model makes precise distinction between the behaviour of the decision maker and the analysis of the researchers. It assumes that the decision-makers have a perfect discrimination capability; however, the researcher does not have complete information about all the elements considered by the individual making a choice. Therefore, the utility $U_{a,it}$, which individual i associates with alternative a in year t, can be decomposed into two parts[27]:

$$U_{a,it} = V_{a,it} + \varepsilon_{a,it} \qquad (4)$$

where $V_{a,it}$ is the deterministic and observable part, which is a function of the measured attributes; and $\varepsilon_{a,it}$ is the stochastic part, capturing the uncertainty, which reflects unobserved alternative attributes, unobserved taste variation and measurement errors made by the researcher.

In the pseudo panel setting, the deterministic part of the utility of the decision-maker ($V_{a,it}$) can be further decomposed into three terms: $V_{a,it} = \overline{V}_{a,ct} + \eta_{a,ct} + \theta_{a,it}, \forall i \in c$, among which only the first component is observable. After the individuals are aggregated into cohorts, the researcher can only observe the average deterministic utility of all sampled individuals in cohort c in year t, $\overline{V}_{a,ct}$; on the other hand, measurement error of true mean utility for the cohort ($\eta_{a,ct}$), and the deviation from the cohort mean utility ($\theta_{a,it}$) are unobservable to the researcher. Furthermore, we assume that the random term $\varepsilon_{a,it}$ has a component of variance structure: $\varepsilon_{a,it} = \lambda_{a,c} + \varepsilon'_{a,it}, \forall i \in c$. As a result, expression (4), the utility of individual i in cohort c year t choosing alternative a, can then be rewritten as:

$$U_{a,it} = \overline{V}_{a,ct} + \eta_{a,ct} + \theta_{a,it} + \lambda_{a,c} + \varepsilon'_{a,it} \qquad (5)$$

[27] We directly start from a panel data model and introduce a time dimension accordingly.

where:

$\overline{V}_{a,ct} = \frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} V_{a,it}, i \in c$, which is the sample mean observable utility of alternative *a* for cohort *c* in year *t*. Note that n_{ct} is the sample size of cohort *c* in year *t*;

$\eta_{a,ct} = \overline{V}^{*}_{a,ct} - \overline{V}_{a,ct}$, representing the measurement error. It is the difference between the sample mean utility and the (unobservable) true mean utility of alternative *a* for cohort *c* in year *t* ($\overline{V}^{*}_{a,ct} = \frac{1}{N_c} \sum_{i=1}^{N_c} V_{a,it}, i \in c$)[28];

$\theta_{a,it}$ represents the unobserved utility of alternative *a* for individual *i* in year *t*, which is the deviation from the mean utility for the cohort. Ignoring measurement errors then $\theta_{a,it}$ is observable to researchers in the cross-sectional models and is "lost" in the aggregation process of pseudo panel;

$\lambda_{a,c}$ is the (time-invariant) unobserved heterogeneity, which includes alternative specific constants and cohort fixed (random) effects. It is assumed to be distributed independently of $\varepsilon'_{a,it}$;

Finally, the last term in (5), $\varepsilon'_{a,it}$ accounts for the randomness besides heterogeneity, which we assume to be independently identically distributed with mean zero and variance σ^2.

The derivation of choice probability for the discrete choice model of pseudo panel will be similar to the standard random utility model. The probability of individual *i* choosing alternative *a* is equivalent to the probability that the utility of alternative *a* is higher than that of any other alternatives:

$$P_{a,it} = \text{Prob}(U_{a,it} > U_{b,it}), \forall b \in A, b \neq a \qquad (6)$$

Substituting (5) into (6), we have:

$$P_{a,it} = \Pr ob(\overline{V}_{a,ct} + \eta_{a,ct} + \theta_{a,it} + \lambda_{a,c} + \varepsilon'_{a,it} > \overline{V}_{b,ct} + \eta_{b,ct} + \theta_{b,it} + \lambda_{b,c} + \varepsilon'_{b,it}), \forall b \in A, b \neq a \qquad (7)$$

---

[28] Note that while cohort sample changes year by year, the cohort population remains fixed over time if cohorts are defined based on time-invariant variables and we assume total population is close, i.e. there is no birth or death and cohort size N_c remains constant over time.

Equation (7) is a general probability model of discrete choice pseudo panel, and its estimation remains a problem. The first issue to confront is the measurement error problem with pseudo panel. For linear model, various error-in-variable estimators have been proposed in the literature. However, their application in empirical work is quite complex and such problem is likely to be greater for non-linear model. As a result, here we only consider (7) under the most common asymptotic in empirical studies: $n_{ct} \to \infty$, i.e. the sample size is sufficiently large for each cohort. In this case, the measurement error converges in probability to zero:

$$\underset{n_{ct}\to\infty}{\text{Plim}}\,\eta_{a,ct} = 0$$

As in the case of linear pseudo panel, the measurement error is asymptotic zero when the number of sample observations n_{ct} is sufficiently large for each cohort c in year t. This condition is likely to be met in the current study, and by ignoring the measurement error $\eta_{a,ct}$, the dimension of integral required to solve (7) is reduced.

Second, we aggregate the two random terms $\theta_{a,it}$ and $\varepsilon'_{a,it}$, and consider their sum as a new composite random variable: $e_{a,it} = \theta_{a,it} + \varepsilon'_{a,it}$. These two random terms are both unobservable to the researcher and while they represent different source of randomness, they are empirically indistinguishable in the pseudo panel setting. Consequently, these two sources of uncertainty have to be combined, which further reduces the dimension of integral required to solve (7). However, it follows that the stochastic part of utility is different between discrete choice models based on pseudo panel and those based on individual data, and hence the parameters estimates are not directly comparable between them.

At this stage, it is not possible to make a general assumption about the unobserved heterogeneity $\lambda_{a,c}$, as the appropriateness of such assumption will depend on the relationship between $\lambda_{a,c}$ and the observed deterministic utility component $\overline{V}_{a,ct}$. Conditional on the unobserved heterogeneity $\lambda_{a,c}$, the cumulative distribution of the aggregated random term $e_{a,it}$ would determine the probability model ($\lambda_{a,c}$ need to be

integrated out, estimated, or tackled using some semi-parametric techniques at a later stage). Rearranging the terms in (7), the probability function can be expressed as:

$$P_{a,it} = \Pr ob[e_{b,it} < e_{a,it} + (\overline{V}_{a,ct} + \lambda_{a,c}) - (\overline{V}_{b,ct} + \lambda_{b,c})], \forall b \in A, b \neq a \quad (8)$$

Note the distribution of the residuals e by $f(e) = f(e_1,...,e_n)$ expression (8) can be written more precisely as:

$$P_{a,it} = \int_{R_n} f(e)de \quad (9)$$

where $R_n = \begin{cases} e_{b,it} < e_{a,it} + \overline{V}_{a,ct} + \lambda_{a,c} - (\overline{V}_{b,ct} + \lambda_{b,c}), \forall b \in A, b \neq a \\ \overline{V}_{b,ct} + \lambda_{b,c} + e_{b,it} \geq 0 \end{cases}$

By assuming specific distributions of *e*, it is possible to derive analytical expressions for model (9) and generate models of different functional forms. What is the most appropriate assumption on the distribution of the random term will certainly vary between studies, so the following section will discuss it in the context of car ownership model.

6.2.2 A Discrete Choice Model of Household Car Ownership

To formulate a Random Utility model of car ownership, the first step is to determine the decision maker and choice set. In the current study, the decision makers are the household *i* within cohort *c*. For a car ownership model, the complete choice set is the number of car owned: 0 car, 1 car… *n* Cars. Due to smaller sample size for households with 3 or more cars, we limit the choice set of our car ownership model to 0 car, 1 car and 2+ cars.

As mentioned above, the different assumptions on the distribution of random residuals give rise to discrete choice model of different forms. The most common model is multinomial logit model (MNL), which assumes the random terms ($e_{a,it}$ in the current study) are distributed IID (independently and identically distributed) Gumbel. Its popularity lies in the low computation costs of parameter estimation. However, it exhibits the Independence of Irrelevant Alternatives (IIA) property, i.e. the ratios of any two probabilities is necessarily the same no matter what other alternatives are in the choice set or what the characteristics of other alternatives are. This property is clearly inappropriate in certain situation, and the MNL has the danger of failure in the

presence of correlated alternatives (e.g. the red bus, blue bus problem). As the two options of owning one car and two plus cars are correlated, MNL might not be appropriate for the car ownership model either. Nevertheless, such structure is popular in the empirical work, especially as part of the bigger modelling system (e.g. Train, 1986; HCG, 2000).

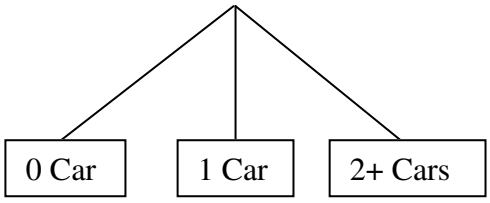
Multivariate probit model relaxes the IID assumption of the random components of the utility. Instead, the random residuals are assumed to be distributed jointly normal, with a general variance-covariance matrix. Because of the assumption on the random utility components is completely general, probit model successfully tackles two problems confronting MNL model: non-independence from irrelevant alternatives as well as taste variation among individuals. The choice probabilities of the probit model are quite complex (multiple integrals with no closed form), and it can only be estimated using alternative approaches such as Clark approximation and Monte Carlo Simulation (Daganzo, 1979; Train, 2003). The application of probit models to the nonlinear pseudo panel model of (9) is even more complicated. More specifically, there are two sources of unobserved heterogeneity, one choice specific and one cohort specific. To distinguish these two sources of heterogeneity, while maintaining the flexible correlation structure of the composite random term $e_{a,it}$, would make consistent estimation of the model extremely difficult.

MNL and probit are "flat" models, which can be illustrated by graph (a) in Figure 6-1. Given the drawbacks of MNL and complexity of probit, a hierarchical model structure becomes an attractive alternative, which is illustrated by graph (b). It involves estimating two binary choice models in steps.

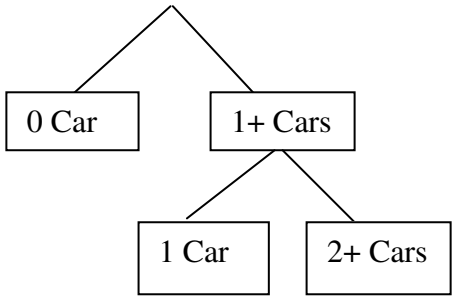
The hierarchical model of car ownership involves two binary choice models: the first is the choice between zero car and one plus cars (noted as Model 1+ hereafter); then conditional on owning at least one car, choice between owning exactly one car and two plus car (noted as Model 2+|1+ hereafter). It should be noted that the hierarchical model of (b) is not a standard Nested Logit Model, which would have the same complication as the multivariate probit model; instead, it consists of two separate binary choice models. For each binary choice model, it does not require the IIA

Figure 6-1 Two Structures of multiple car ownership modelling

(a) "Flat" model of MNL and probit



(b) Hierarchical model structure



assumption and the assumption on the random term can be general. Moreover, such formulation facilitates the consistent estimation of the unobserved heterogeneity and dynamic effect, thanks to a growing econometric literature in this field. It also has the advantage of choice probability (of higher car ownership) increases monotonically with income. As a result, the hierarchical model structure is adopted for the current project, similar to other car ownership models such as NRTF (1997), Whelan (2001) and RAC (2002b).

After determining the decision maker, choice set and model structure, the empirical model of car ownership can be readily identified. For Model 1+, the utility of owning no car is normalized to zero: $U_0 = 0$; on the other hand, the utility of owning at least one car can be expressed as a linear function of:

$$U_{1+} = \bar{x}'_{ct}\beta + \lambda_c + e_{it} \qquad (10)$$

where $\bar{x}'_{ct}$ is a vector of explanatory variables for cohort *c* in year *t*, including car purchase price, car running costs, income and other relevant household characteristics[29]. While all the households in cohort *c* have the same mean deterministic utility ($\bar{x}'_{ct}\beta$) and unobserved cohort heterogeneity (λ_c), they have different composite error term e_{it}. This reflects the essence of the Random Utility Model: given the same observed deterministic utility, decision makers behave differently due to the unobserved random

[29] For dynamic models, lagged dependent variable $\bar{y}_{c,t-1}$ could also be included, although it could impact the consistent estimation of the model. These issues will be discussed in the next chapter.

error. In the current study, this is manifested in the fact that only a proportion of households in a cohort choose to own car(s). Note the household in cohort *c* owning at least one car in year *t* is noted as $y_{ct}^{1+}=1$, then:

$$y_{ct}^{1+}=1 \Leftrightarrow U_{1+}>U_0 \Leftrightarrow \bar{x}_{ct}'\beta+\lambda_c+e_{it}>0 \qquad (11)$$

Assuming the distribution of e_{it} is IID logistic, the probability of household *i* in cohort *c* owning at least one car is that of a familiar logit model:

$$P_{1+}=\Pr ob(y_{ct}^{1+}=1 \mid \bar{x}_{ct}', \lambda_c)=\frac{\exp(\bar{x}_{ct}'\beta+\lambda_c)}{1+\exp(\bar{x}_{ct}'\beta+\lambda_c)} \qquad (12)$$

If e_{it} is assumed to follow a normal distribution, the car ownership model becomes a probit model. As the normal and logistic distributions are similar except in the tails, these two models should give similar predictions except that one choice completely dominates the other. The choice of functional form is sometimes determined for practical reasons, and will be investigated in the empirical section later this chapter.

Model 2+|1+ would be estimated using a reduced pseudo panel dataset of car owning households, while having the identical formulation of Model 1+. The utility of owning exactly one car will be normalized to zero, and the utility of owning two or more cars will be defined in a similar fashion. The choice probability of household owning two or more cars conditional on owning at least one car ($P_{2+|1+}$) will also be similar to (12) based on the IID logistic assumption of the composite random term.

Finally, we discuss the interpretation of the linear utility of owning a car, i.e. expression (10). Households derive utility of car ownership from driving the car; as a result, it seems odd that income, price and other household characteristics are appropriate explanatory variables to be included in the utility function. This is a recognized issue and in the literature of joint car ownership/use model, the utility function of (10) is interpreted as the linear approximation of the conditional indirect utility function. The conditional indirect utility function is the function that gives maximum utility achievable at given prices and income, conditional on the choice of a certain alternative. As shown by Varian (1992), consumers' preferences can be equivalently represented by a direct utility function or an indirect utility function. Starting with the latter, it is relatively straightforward to derive the demand function of

car use using Roy's identity[30] (Train, 1986). Although the current study deals with car ownership only, it is more appropriate to interpret (10) as the (linear approximation of) conditional indirect utility function, and the extension to car use would also become straightforward.

6.3 Estimation of Discrete Choice Pseudo Panel Model

In the previous section, we formulate the random utility model of pseudo panel under the asymptotic of $n_{ct} \to \infty$, ignoring the measurement error problem. In this case, it can be treated as genuine panel and estimated using similar techniques (although it should be weighted by the number of observations in the cohort sample, as it is shown next). However, the consistent estimation of discrete choice model with (genuine) panel data is non-trivial and the most familiar fixed effect model suffers the incidental parameter problem when T is small. Random Effect models can be consistently estimated using maximum likelihood methods, although the likelihood function involves an integral with no closed form so Gauss-Hermite quadrature or simulation method has to be used. The drawback of the random effect model is that it makes restrictive assumption on the distribution of the unobserved heterogeneity and relies on the orthogonality assumption between the explanatory variables and the unobserved effect.

A growing literature has explored various semi-parametric models, without specifying a parametric form of the distribution of the underlying errors, or the distribution of the individual effects conditional on the explanatory variables. While semi-parametric models have growing significance as analytical models, their limitations are the inability to calculate marginal effects based on these estimators and the need for larger samples. As a result, semi-parametric models will only be mentioned as passing reference. Finally, we limit our discussion to the binary choice model, although some of the methods can be extended to the multinomial case readily.

[30] An alternative method (De Jong, 1989a) is to start from the empirical demand function of car use and derive the indirect utility function, also using Roy's identify.

6.3.1 Fixed Effect model

For panels and pseudo panels, fixed effect models do not assume specific distribution of the unobserved heterogeneity λ_c. This unrestrictive feature makes fixed effect model particularly attractive. However, unlike the linear model, whose fixed effect can be eliminated by demeaning or differencing, it is not possible for the case of nonlinear panel model. Nevertheless, in the case of a limited number of individuals observed over many time periods, one can justify treating the cohort fixed effects as parameters to be estimated (Honore, 2002). More specifically, the maximum likelihood estimation of panel data fixed effect model is consistent when $T \to \infty$. For pseudo panel, if the measurement error problem can be ignored, FE model can also be estimated by maximum likelihood using cohort dummy variables.

However, when estimating discrete choice pseudo panel data, it is important that the data are weighted by the number of the observations in each cohort. For binary choice model, noting the proportion of decision makers in cohort *c* making Choice 1 in year *t* ($y_{it} = 1, \forall i \in c$) as r_{ct}, un-weighted MLE assumes that r_{ct} is from a distribution with variance $r_{ct}(1-r_{ct})$. However, the unconditional variance is in fact $r_{ct}(1-r_{ct})/n_{ct}$, where n_{ct} is the number of observations in the cohort sample in year *t*, so the efficiency of maximum likelihood estimation using un-weighted data is underestimated (Greene, 1995). Furthermore, if the choice proportion r_{ct} is based on different numbers of observations, the variances will differ correspondingly, so the un-weighted model will not account for the inherent heteroskedasticity of the pseudo panel model. The maximum likelihood estimation of standard logit or probit model based on proportions data (the form of pseudo panel data), both weighted and unweighted, has been implemented in standard econometric software such as Limdep[31].

The above discussion will become clear with the derivation of log likelihood function for the pseudo panel data. Note the number of individuals in the cohort sample making Choice 1 in year *t* as m_{ct}, we have $m_{ct} = n_{ct} \cdot r_{ct}$, with r_{ct} and n_{ct} as defined above. The likelihood function can be expressed as:

[31] However, random effect model is not available for models with proportions data.

$$L = \prod_{c=1}^{C}\prod_{t=1}^{T}(P_{ct})^{m_{ct}}(1-P_{ct})^{n_{ct}-m_{ct}} = \prod_{c=1}^{C}\prod_{t=1}^{T}[(P_{ct})^{r_{ct}}(1-P_{ct})^{1-r_{ct}}]^{n_{ct}} \qquad (13)$$

where P_{ct} is the probability of the individuals in cohort c making Choice 1 in year t. In the current study of car ownership, it could take the form of equation (12).

Taking logarithm of expression (13) we have derived the log likelihood function of logit model based on proportions (pseudo panel) data:

$$\ln(L) = \sum_{c=1}^{C}\sum_{t=1}^{T} n_{ct}[r_{ct}\ln(P_{ct}) + (1-r_{ct})\ln(1-P_{ct})] \qquad (14)$$

Comparing the log likelihood function of (14) with that of binary choice model based on individual (discrete) data, it is clear that the only difference is the introduction of weighting n_{ct}. It should be noted that while the discussion here is based on binary choice model, it can be easily extended to the case of multinomial choice model.

However, the maximum likelihood estimator of the Fixed Effect model is only consistent when the number of time period is large, i.e. $T \to \infty$. This can be illustrated using the analysis of asymptotic variance. Rewrite the log likelihood function of (14) as:

$$Ln(L) = \sum_{c=1}^{C}\sum_{t=1}^{T} Ln[g(r_{ct}, n_{ct}, \beta' x_{ct} + \lambda_c) \qquad (15)$$

If β were known, the solution for λ_c would be based on only the $T(c)$ observations for cohort c. This implies that the asymptotic variance for λ_c is of order $T(c)$. Because β is not known, is has to be estimated, and the estimator is a function of the maximum log likelihood estimator of λ_c. As a result, the asymptotic variance of b_{ML} (maximum log likelihood estimator of β) must also be of order $T(c)$. In another word, the MLE of β is a function of a random variable which does not converge to a constant as $C \to \infty$ (Greene, 2001a; 2001b). This is the incidental parameter problems as identified in Neyman and Scott (1948). This problem can also be explained intuitively. For nonlinear panel model in general (and same for nonlinear pseudo panel), the incidental parameter λ_i can not be differenced away as in the case of linear model. Only new observations for individual i give new information about λ_i; however, given a fixed T more individuals do not help with the estimation of λ_i because they add more parameters to be estimated.

In some cases, the pseudo panel might be constructed in a way that the number of cohorts is large and the number of observations per cohort is small. If the number of time periods is also small, it would be appealing to consider asymptotic with large C and fixed T. However, further research is required to establish the consistent pseudo panel estimator under such asymptotic, as measurement error needs to be taken into account. For genuine panel, fixed T consistent estimator has been proposed in the literature. For example, although the nuisance parameter λ_i can not be differenced away, $\sum y_{it}$ is the sufficient statistic for binary logit and other small class of models. Conditional on the sufficient statistic $\sum y_{it}$, conditional maximum likelihood estimator would produce unbiased estimate of β (Andersen, 1970). The conditional maximum likelihood estimator is extended to the multinomial logit by Chamberlain (1980).

While $\sum y_{it}$ is the sufficient statistic for panel data logit model, the existence and form of such statistic are difficult to establish for the case of pseudo panel. Furthermore, the conditional maximum likelihood approach has one significant drawback: it does not allow the calculation of the average effect of x_{it} on the probability of $y_{it} = 1$ across the distribution of λ_i. The similar problem applies to other semi-parametric estimators such as the maximum score estimator of Manski (1987).

Another class of estimators does not attempt to be fixed T consistent; instead, the objective is to reduce the biases rather than to eliminate biases completely. One prime example is the modified concentrated likelihood estimator proposed by Arellano (2003), which has bias of order $1/T^2$ rather than the maximum likelihood estimator of $1/T$. The modified concentrated likelihood estimation has been extended to dynamic panel by Carro (2003), and both will be discussed in the next chapter.

Given that the maximum likelihood estimator of the fixed effect model is not consistent, it is important to establish the extent of biases. The discrete choice fixed effect estimator shows a substantial finite sample bias when the number of time period is very small. Hsiao (1986) found that for $T = 2$, the maximum likelihood estimator for a binary logit model is 100%. However, such large bias might only be of theoretical

importance, as the bias reduces rapidly as T increases to 3 or more. In another widely cited study, Heckman (1981b) found that the small sample bias of fixed effect estimator is surprisingly small even with moderate T. Using Monte Carlo simulation, the author showed that for the probit model with sample size of $T_i = 8$ and $N = 100$, the bias of the slope estimator is on the order of only 10%. In a more recent study, Greene (2002) found that the bias in the marginal effects is smaller than the bias in the slope parameters of MLE, which suggests that even when $T = 2$, the bias of 100% might also be overstated.

In the current study, 11 out of a total of 16 cohorts have pseudo panel observations for 19 years. As a result, the problem of small T bias for the MLE of fixed effect model might not be significant. In the empirical study of car ownership, the use of weighted MLE for the discrete choice pseudo panel model with fixed effects seems justified.

6.3.2 Random Effect Estimators

The "classic" random effect estimator assumes that the unobserved heterogeneity λ_c is unrelated to the explanatory variables X, so that the conditional distribution $f(\lambda_c \mid X)$ is not dependent on X. This is an assumption that is restrictive and in many case unrealistic, so a number of studies attempt to develop estimators between random effect and fixed effect, i.e. not requiring the orthogonality assumption while un-affected by the incidental parameter problem. In this section, we first briefly discuss the estimator of "classic" random effect model; then we introduce some "generalized" random effect estimators that allow correlation between the unobserved heterogeneity and the explanatory variables.

A random effect discrete choice model was first implemented by Bulter and Moffitt (1982) and was subsequently implemented in some econometric software such as Limdep (Greene, 1995). It should be noted that random effect model in Limdep does not apply to proportions data, so the random effect pseudo panel model can not be estimated using Limdep. Nevertheless, there is no reason why this approach can not be extended to the pseudo panel case if the assumptions on the error terms can be maintained.

In a binary choice model with error terms of $u_{it} = \varepsilon_{it} + \lambda_i$, assuming the two error components are independent random variables with

$E[\varepsilon_{it} \mid X] = 0$; $Cov[\varepsilon_{it}, \varepsilon_{js} \mid X] = Var[\varepsilon_{it} \mid X] = 1$ if $i=j$ and $t=s$; 0 otherwise

$E[\lambda_i \mid X] = 0$; $Cov[\lambda_i, \lambda_j \mid X] = Var[\lambda_i \mid X] = \sigma^2$ if $i=j$; 0 otherwise

$Cov[\varepsilon_{it}, \lambda_j \mid X] = 0$ for all i, t, j.

Based on the above assumption, u_{it} are independent conditional on the unobserved heterogeneity λ_i. As a result, by integrating λ_i out of the joint density of $(u_{i1},...,u_{iT},\lambda_i)$, the likelihood function of joint probability of all T_i observations can be simplified to an one dimensional integral of:

$$L_i = P[y_{it},..., y_{iT} \mid X] = \int_{-\infty}^{+\infty} [\prod_{t=1}^{T} \Pr ob(Y_{it} = y_{it} \mid x_{it}'\beta + \lambda_i)] f(\lambda_i) d\lambda_i \qquad (16)$$

The inner probability of (16) can be any forms of discrete choice model including probit, logit, etc. The question remains how to do the outer integration. Basically, there are two approaches. The first one relies on the normality assumption of λ_i, and after a bit of manipulation, (16) can be further reduced to a function that is amenable to Gauss-Hermite quadrature for computation. This is the approach used in Limdep and other modern econometric software. The second approach is the method of maximum simulated likelihood. It allows more flexibility in the distribution of λ_i, and it is straightforward to extend the random effect model to random parameter model and to the case of pseudo panel by weighting the log likelihood function. The maximum simulated likelihood method will be discussed in the next chapter.

One appealing feature of the fixed effect model is that the unobserved heterogeneity λ_i is allowed to correlate to the explanatory variables X. A number of authors have tried to achieve this in a random effect model by parameterizing the distribution of λ_i as a function of x_i. The most notable example is Chamberlain (1984), in which the following assumption is made:

$$\lambda_i \mid (x_{i1},..., x_{iT_i}) \sim N(\sum_{t=1}^{T_i} x_{it}'\gamma_t, \sigma_t^2) \qquad (17)$$

where the parameters γ_t and σ_t^2 might depend on T_i. For a probit model, after necessary normalizations for identification, $\{\gamma_t\}_{t=1}^{T}$, as well as the slope coefficient β, can be estimated using maximum likelihood.

Newey (1994) and Chen (1998) are the semi-parametric extension of Chamberlain (1984), where the conditional mean is assumed to be $\rho(x_{i1},...,x_{iT})$ with the function ρ remains unspecified. These models are reviewed in Arellano and Honore (2001) so they will not be repeated here.

As pointed out by Honore (2002), for (17) to hold for all T_i, it would place strong assumptions on the distribution of the explanatory variables. As implied by the law of iterated expectations, if (17) holds in both time periods T and T+1, then the conditional mean of x_{iT+1} would be a linear function of its past value, whose coefficients are directly related to those of a probit model for the distribution of y_{it}. This assumption makes the "generalized" random effect model much less appealing then it first appears, so it will not be investigated in the empirical work in the current study.

6.4 Empirical Results of Static Car Ownership Model

In this section, the results of static car ownership model will be reported. Investigation is made on the forms of the explanatory variables (average household size and other demographic characteristics, splits of household types, log transformation, etc.), functional form (logit and probit), fixed effect and random effect. Separate results will be reported for models of owning at least one car, and those of owning two or more cars conditional on owning the first one.

Both Limdep and Gauss have been used in the estimation. The Gauss code used in the current project is specially adapted for pseudo panel data and has been used to estimate most of the models in the current study. Limdep is used for some specific models not implemented in the Gauss code, such as probit model, and is also used as validation of the correct implementation of the Gauss code.

6.4.1 Models of One Plus Cars

The pseudo panel dataset used for the models of one plus cars is the same as that used in the linear model of Chapter 4 and 5. It has 254 observations, covering 16 cohorts from years 1982 to 2000 (not all cohorts have observations for all periods). The descriptive statistics of the data was discussed in the previous chapters so is not repeated here.

Systematic specification search has been conducted to determine the model with the best fit. As car ownership is influenced not only by income and price, but also by household structure (demographic characteristics) and the proxy for accessibility, location, all these variables should be included in the indirect utility function of car ownership. Household demographic characteristics include average number of children per household in a cohort, average number of person in work and average household size, which can be directly used as explanatory variables. Alternatively, there is an eight-way categorization of the household types based on these three variables (for detailed description see Table 3-2 Chapter 3), and the split of each household types within a cohort can be used as the explanatory variables.

The household locations are divided into five categories including Greater London, metropolitan areas and other areas with varied population density. Proportion of households within a cohort living in each of the area types (dropping one for identification) can be used as explanatory variables. Alternatively, proportions of households living in Greater London and metropolitan areas are combined in a variable "MET", which is then included in the utility function, together with the proportion of households living in the least populated rural areas (re-named from "Area5" to "Rural" for clarity). Table 6-2 reports the logit model results with different representation of household characteristics and locations.

Besides the household characteristics and location variables, the variables common to model 1-4 are: Constant (ONE), average weekly real disposable income per household (Inc), index of real car purchase price (Price), index of real car running costs (RunCst), average age of the household head in the cohort (Age), and square of cohort age divided by 100 (AgSq). Table 6-2 shows the model coefficients with the t-statistics in the parenthesis. Across all four models, slope coefficients for income are always

Table 6-2 Logit Model 1-4, alternative variables for household characteristics and location (t-stat in the parenthesis)

	Model 1		Model 2		Model 3		Model 4	
ONE	-0.7100	(-2.34)	-0.7477	(-1.59)	-0.4506	(-0.76)	0.3870	(0.81)
Inc	0.0028	(8.13)	0.0027	(7.84)	0.0023	(6.94)	0.0023	(7.06)
Child	0.2067	(1.78)	0.1474	(1.24)				
Worker	-0.1643	(-2.35)	-0.1868	(-2.65)				
HHSize	-0.0718	(-0.55)	-0.0130	(-0.10)				
HH2					-0.8050	(-1.75)	-0.7243	(-1.59)
HH3					-1.6564	(-2.94)	-1.6361	(-2.93)
HH4					1.2817	(3.12)	1.3235	(3.26)
HH5					1.1509	(2.42)	1.1888	(2.51)
HH6					1.2524	(3.21)	1.3083	(3.40)
HH7					-0.1193	(-0.24)	-0.0579	(-0.12)
HH8					0.5463	(1.20)	0.6118	(1.35)
Area2			-0.9704	(-1.87)	0.1482	(0.28)		
Area3			-0.2651	(-0.53)	0.6459	(1.25)		
Area4			0.5199	(1.10)	1.2575	(2.58)		
Area5			1.5368	(3.38)	1.6069	(3.45)		
Met	-0.7775	(-2.72)					-0.8340	(-2.89)
Rural	1.3056	(4.23)					0.6323	(2.02)
Price	-0.0077	(-4.67)	-0.0086	(-5.11)	-0.0170	(-9.38)	-0.0166	(-9.30)
RunCst	-0.0090	(-6.65)	-0.0091	(-6.65)	-0.0068	(-4.84)	-0.0071	(-5.04)
Age	0.1309	(19.17)	0.1288	(18.71)	0.0848	(10.79)	0.0866	(11.10)
AgSq	-0.1416	(-21.77)	-0.1402	(-21.47)	-0.0905	(-10.89)	-0.0922	(-11.17)
LL	-70047		-70044		-69938		-69940	
Null LL	-85914		-85914		-85914		-85914	
Adj. LRI	0.1845		0.1846		0.1857		0.1857	

LL: Log Likelihood;
Null LL: Null Log Likelihood;
Adj. LRI: Adjusted Likelihood Ratio Index (sometimes called Rho bar square); it is calculated as 1 - (LL - K) / Null_LL, where K is the number of explanatory variables.

positive, significant and of very similar magnitude. The slope coefficients for car purchase price index and running costs index are always negative and significant. The coefficients for cohort age and age squares are highly significant, suggesting a strong life cycle effects of car ownership. The coefficients for cohort age are always positive, while those for cohort age square are always negative, indicating a peak of car ownership over the life cycle.

Model 1 and 2 include the average number of children, person in work and household size as explanatory variables. However, the coefficients for "Child" and "HHSize" are not statistically significant, and those for "Worker" have the unexpected negative sign. The statistics of model fit are reported in the bottom part of Table 6-2, and Model 1 and 2 have smaller log likelihood compared to Model 3 and 4 with lower Adjusted

Likelihood Ratio Index at about 0.1845. This shows that Model 1 and 2 have the worse level of fit than the other two models, where the eight-way categorization is used to describe household demographic characteristics.

In Model 3 and 4, the proportion of household type 1 (single working adult household) is dropped, so the coefficients for other household types should be interpreted in relation to type 1. The coefficients for type 2 and 3, both being single adult household, are negative, indicating lower propensity of car ownership for these two household types compared to the base case of type 1. The coefficients for type 4 to 8 (Household with two ore more adults) are mostly positive (one negative coefficient is not statistically significant at all), indicating higher propensity of car ownership. Regarding the location variables, Model 3 includes four location types while in Model 4 the location types are compressed into two. However, the two models have very similar log likelihood; the Likelihood Ratio Test produces a chi square statistic of 3.18 with 2 degree of freedom, which is not significant at 10% level. This result shows that there is no loss of fit in compressing the area type, so model 4 should be regarded as the preferred model.

Table 6-2 reports the slope coefficients, which are different from the marginal effects for discrete choice models. Marginal effect measures the impacts of a small change of explanatory variable on the choice probability, and for logit model it is calculated by[32]:

$$\frac{\partial E(y \mid X)}{\partial x} = \frac{\exp(x'\beta)}{[1+\exp(x'\beta)]^2} \cdot \beta \qquad (18)$$

which depends not only on the slope coefficients, but also on the value of explanatory variable. The marginal effects can be evaluated at the sample means of the data; alternatively, one can evaluate the marginal effects at each observation then calculate the average effect over the sample observations. Table 6-3 reports the marginal effects based on the first method, evaluating at the weighted average of the explanatory variables (weight being the number of sample households within a cohort).

[32] Expression (18) calculates the marginal effect for continuous variable, although it also provides "an approximation that is often surprisingly accurate" for dummy variables (Greene, 2003, p668).

Table 6-3 Marginal Effect at weighted average of explanatory variables, Model 1 – 4

	Model 1	Model 2	Model 3	Model 4
ONE	-0.14300 **	-0.15059 '	-0.09081 '	0.07800 '
Inc	0.00056 ***	0.00054 ***	0.00046 ***	0.00046 ***
Child	0.04163 *	0.02968 '		
Worker	-0.03308 **	-0.03762 ***		
HHSize	-0.01445 '	-0.00262 '		
HH2			-0.16223 *	-0.14597 '
HH3			-0.33381 ***	-0.32972 ***
HH4			0.25829 ***	0.26671 ***
HH5			0.23194 **	0.23957 **
HH6			0.25240 ***	0.26365 ***
HH7			-0.02403 '	-0.01166 '
HH8			0.11010 '	0.12328 '
Area2		-0.19544 *	0.02986 '	
Area3		-0.05338 '	0.13017 '	
Area4		0.10471 '	0.25342 ***	
Area5		0.30952 ***	0.32383 ***	
Met	-0.15659 ***			-0.16808 ***
Rural	0.26296 ***			0.12743 **
Price	-0.00155 ***	-0.00173 ***	-0.00342 ***	-0.00334 ***
RunCst	-0.00182 ***	-0.00184 ***	-0.00138 ***	-0.00142 ***
Age	0.02637 ***	0.02594 ***	0.01709 ***	0.01744 ***
AgSq	-0.02852 ***	-0.02823 ***	-0.01823 ***	-0.01857 ***

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Besides the marginal effects, sometimes it is useful to interpret the results in terms of elasticity, which can be calculated by $El = (\partial P/\partial x)(x/p)$ when the explanatory variable x is in linear form (no logarithm or other transformation). Note that the first term $(\partial P/\partial x)$ is the marginal effect as expression (18), which depends on the evaluated values of all explanatory variables. Table 6-4 reports the income elasticity, purchase price elasticity and running costs elasticity for Model 1 and Model 4.

Table 6-4 Income, price and running costs elasticity for household with various income

	Model 1			Model 4		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.29	-0.49	-0.51	0.24	-1.04	-0.40
Median	0.24	-0.19	-0.29	0.18	-0.38	-0.20
High	0.21	-0.14	-0.14	0.17	-0.30	-0.11

Low income: 10 percentile of income, £172 per week;
Mid income: 50 percentile of income, £306 per week;
High income: 90 percentile of income, £430 per week.

Both models show the relatively low income elasticity, and the difference between the low income household and high income household is very small. Based on model 4, the results suggest that a 1% increase of income would increase of probability of owning at least one car by 0.24% from 0.3975 to 0.3984 for low income household; for high income household, such probability is increased by 0.17% from 0.8267 to 0.8281. For those with high income, the proportion of households owning at least one car is already very high, so it is reasonable to expect low income elasticity. On the other hand, when income is specified as linear term in the utility function, it implies that a £1 increase has the same impact on utility of car ownership whether the weekly household income is £100 or £1,000. This implication can be problematic, and it can be tackled by the nonlinear transformation of the income variables. The methods of nonlinear transformation include logarithm, Box-Cox, piecewise, power, etc. We will investigate the most common method—logarithm transformation latter this section.

Purchase price elasticity and running costs elasticity, on the other hand, are higher than the income elasticity, especially for low income household. Although some earlier studies (e.g. Dargay and Vythoulkas, 1999; Whelan 2003) found price elasticity to be lower than income elasticity, results similar to Table 6-4 are reported for the some linear pseudo panel models in Chapter 4 and 5. Based on Model 4 for mid income household, a 1% increase of purchasing price would reduce the probability of owning one or more car by 0.38% from 0.7484 to 0.7455; a similar increase of running costs would reduce the probability of car ownership by 0.2% to 0.7469.

After the investigation on the representation of household demographic characteristics and location variables, the issue of model functional form has been examined. We limit our comparison to the two most common models (logit and probit), as other models such as Weibull and Gompertz are rarely used in empirical work. As logistic and probit distributions mainly differ towards the tails, the use of either model would have significant impacts on prediction if the probability of choosing one alternative is high for most observations. On the other hand, if there are few cases of extremely high or extremely low probability, logit model and probit model would produce similar results (after accounting for different model scale). Figure 6-2 shows the observed probability (equal to the aggregate proportion) of owning at least one car and the predicted probability based on Model 4.

Figure 6-2 Observed (Y) & Predicted (P) probability of household owning 1+ car by income

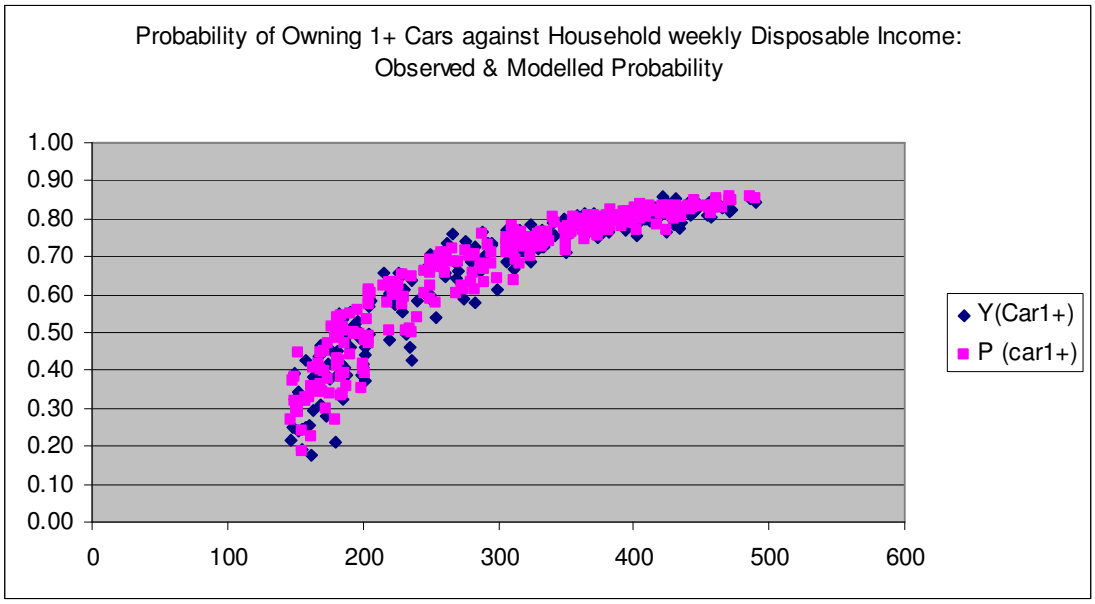


Figure 6-2 does not reveal any systematic difference between the observed and predicted probability, suggesting the appropriateness of the logit functional form. It also illustrates the lack of extreme choice probability, thus more or less eliminating the difference between the logit and probit model. As a matter of fact, the corresponding probit versions of Model 1 to 4 have almost identical marginal effect as Table 6-3.

Next we investigate the problem of heteroskedasticity. Arranging the residual term in Model 4 by average household income, as shown in Figure 6-3, it appears that the variance of residual is larger for the very low income group. To test whether such heteroskedasticity is significant, we estimate a model that has different scale (random errors) for households with low income (weekly income lower than £180) and other households. However, the scale parameter is not statistically significant, suggesting that heteroskedasticity might not be a serious problem.

The next set of test examines the impact of logarithm transformation of income and price variables. The results for models with compressed location variables ("Met" and "Rural") are reported in Table 6-5, with Model 5 and 6 corresponds to Model 1 and Model 4 in Table 6-2 and 6-3.

Figure 6-3 Residual against household income

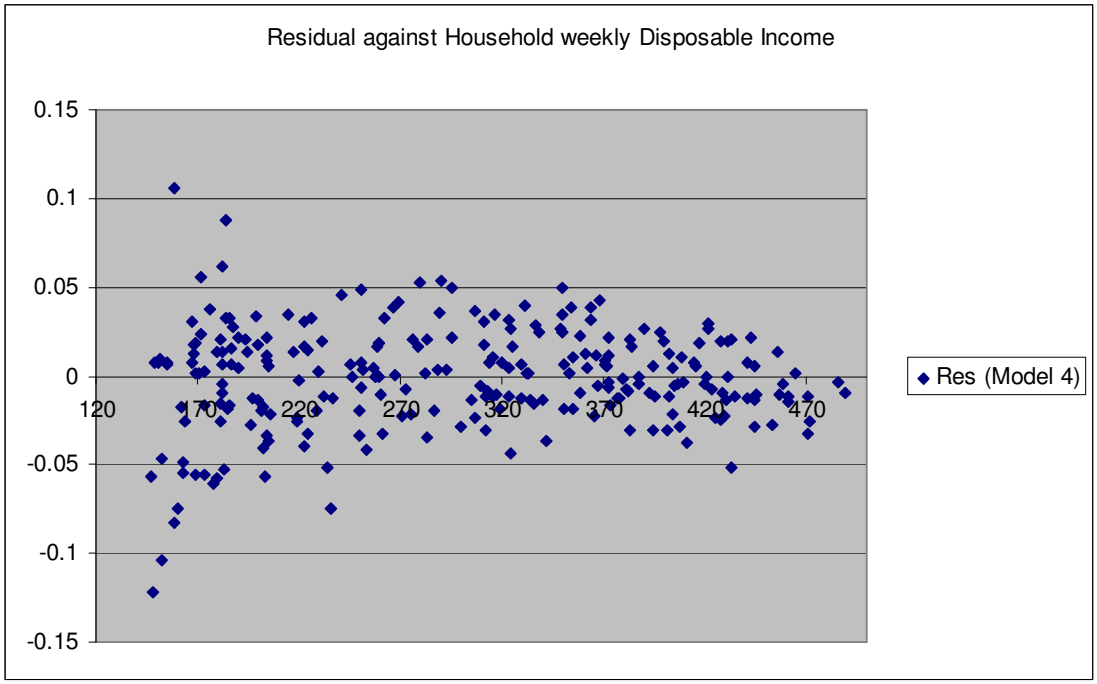


Table 6-5 Models with Log Income and Log price variables (t-stat in parenthesis)

	Slope Coefficient				Marginal Effects			
	Model 5		Model 6		Model 5		Model 6	
ONE	-7.0711	(-4.10)	-0.7656	(-0.40)	-1.4246	***	-0.1543	'
LnInc	1.5463	(14.10)	1.1500	(10.55)	0.3115	***	0.2318	***
Child	-0.0464	(-0.40)			-0.0094	'		
Worker	-0.4357	(-6.24)			-0.0878	***		
HHSize	0.1836	(1.40)			0.0370	'		
HH2			-0.4735	(-1.05)			-0.0955	'
HH3			-1.4396	(-2.59)			-0.2902	***
HH4			1.3115	(3.30)			0.2644	***
HH5			0.6998	(1.49)			0.1411	'
HH6			0.9122	(2.37)			0.1839	**
HH7			-0.1397	(-0.29)			-0.0282	'
HH8			0.1085	(0.24)			0.0219	'
Met	-0.7433	(-2.60)	-0.7996	(-2.77)	-0.1498	***	-0.1612	***
Rural	0.6715	(2.16)	0.3309	(1.06)	0.1353	**	0.0667	'
LnPrice	-0.1578	(-0.92)	-1.0336	(-5.20)	-0.0318	'	-0.2084	***
LnRunCst	-0.3743	(-2.57)	-0.3828	(-2.53)	-0.0754	***	-0.0772	**
Age	0.0893	(11.55)	0.0725	(8.89)	0.0180	***	0.0146	***
AgSq	-0.0983	(-13.03)	-0.0787	(-9.18)	-0.0198	***	-0.0159	***
Log Like'd	-69983		-69921					
Null LL	-85914		-85914					
Adj. LRI	0.1853		0.1860					

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Compared to Model 1, the log likelihood of Model 5 has increased by 65; compared to Model 4, that of Model 6 has increased by 19. As the two corresponding models have the same number of explanatory variables, the higher log likelihood indicates a better fit for models with log income and log price variables. The likelihood ratio test between Model 5 and Model 6 has a Chi square statistic of 123, suggesting a significant loss of fit when average household demographic factors are used instead of proportions of household types. The coefficients and marginal effects of the household characteristics and location variables are broadly similar between Model 4 and Model 6. However, the income and price coefficients and marginal effects are not directly comparable. For Model 5 and 6, the income, purchase price and running cost elasticity (calculated by $El = (\partial P / \partial x) / P$) are reported in Table 6-6.

Table 6-6 Elasticity derived from models with log income and log price variable

	Model 5			**Model 6**		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.95	*-0.10*	-0.23	0.70	-0.63	-0.23
Median	0.42	*-0.04*	-0.10	0.29	-0.26	-0.10
High	0.29	*-0.03*	-0.07	0.21	-0.19	-0.07

(Note: the price elasticity for Model 5 is not statistically significant, hence in italic.)

The income elasticity derived from the log variable model is broadly similar to that derived from the linear variable model for high income households. However, the income elasticity is much higher for low income households, suggesting the rise in income has a much bigger impacts on car ownership for poorer households. The results obtained from linear pseudo panel model (Chapter 4 and 5) showed a similar picture, where the income elasticity derived from semi-log model more than doubles that from the linear model for low income households. Based on Model 6, a 1% income rise would increase the probability of owning 1+ car from 0.3920 to 0.3947 for low income households; from 0.8195 to 0.8212 for those with high income.

The purchase price elasticity and running costs elasticity are lower than the income elasticity, a similar result to the (semi-log) linear pseudo panel model. Both price and running cost elasticity for low income households is about three times of that for high income households. The running costs elasticity derived from Model 5 and Model 6 is almost identical, while the purchase price coefficient is not significant for Model 5 so

we are unable to make any meaningful comparison. Overall, these elasticity estimates look sensible, which should give us some confidence in the estimated model.

Finally, we investigate the fixed effect model and random effect model. The fixed effect models are estimated by adding cohort dummy variables and estimating their coefficients in a similar way as other explanatory variables. This method is justified because the number of time periods is not very short, and maximum likelihood estimator is unbiased when $T \to \infty$. The Fixed Effect models have higher log likelihood. The fixed effect version of Model 4 increases log likelihood by 36, and the LR Test is significant at 1% level (Chi square statistic of 72 with 14 degree of freedom). The Random Effect Model assumes that the unobserved heterogeneity follows a normal distribution, and it is estimated by maximum simulated likelihood method. However, the additional error component is not significant at all and there is no improvement of log likelihood with Random Effect model.

Fixed Effect model and Random Effect model are the most common and important panel (and pseudo panel here) data models and will be further investigated with the dynamic models in the next chapter. As a result, the details of the static FE and RE models are not reported here.

6.4.2 Models of Two plus Cars Conditional on Owning the First Car

As the model of two plus cars is conditional on households owning at least one car, a second pseudo panel dataset (called Dataset 2 hereafter) has been constructed from the sub-sample of car owning households. For the measurement error problem to be ignored, only cohorts with a sufficiently large number (over 100) of observations are included. This reduces the number of pseudo panel observations from 254 of the full dataset to 220 of Dataset 2.

Dataset 2 contains 14 cohorts, with the oldest cohort born between 1906 and 1910 and the youngest cohort born between 1971 and 1975. Two cohorts in the full pseudo panel dataset have to be excluded due to insufficient survey sample size. Table 6-7 presents the descriptive statistics of the pseudo panel Dataset 2.

Table 6-7 Descriptive Statistics of Pseudo Panel Dataset 2

	Car	Inc	Child	Adult	Worker	HHSize	Area2	Area3	Area4	Area5
Median	1.30	369.20	0.42	1.95	1.53	2.53	0.19	0.22	0.23	0.27
Mean	1.31	365.83	0.60	2.02	1.30	2.62	0.19	0.22	0.23	0.27
Stdev	0.16	85.35	0.58	0.21	0.65	0.62	0.03	0.03	0.03	0.04
Max	1.67	534.10	1.76	2.63	2.33	3.84	0.27	0.33	0.31	0.43
Min	1.04	209.03	0.00	1.64	0.08	1.66	0.07	0.12	0.17	0.15

As Dataset 2 is constructed based on car owning households, it is expected that the minimum number of cars per household is greater than one. Compared to the full pseudo panel dataset (summary descriptive statistics in Table 4-1), car owning households have higher income (median weekly real income of £369 compared to £302 in the full sample), bigger household size (median size of 2.53 instead of 2.27) and more persons in work (median of 1.53 instead of 1.27).

Systematic specification search has been conducted for the model of two plus cars. We start from the binary logit model, investigating the representation of household structure and locations variables, while the income and price variables are specified in linear forms. In contrast to the model of one plus car, compressing the location types from four to metropolitan area and least populated rural area leads to significant loss of fit. The Likelihood Ratio Test of compressing location variables produces a Chi square statistic of over 16 with 2 degree of freedom, which is significant at 1% level. On the other hand, whether the average demographic statistics (household size etc.) or proportions of household types are used to represent household characteristics have small impacts on the model fit. Table 6-8 reports the slope coefficients and marginal effects of the two models with detailed location variables (dropping Area1, Greater London for identification).

The marginal effects of the common variables in Model 2 and Model 3 are quite similar. When there is higher proportion of households for any cohort living in areas other than Greater London, the share of households owning two or more cars increases. Interestingly, the propensity of owning 2+ cars is not the highest for households living in the least populated rural area (Area5), which is different from the model of 1+ car. Regarding the household characteristic variables, the higher number of persons in work would increase the probability of owning 2+ cars. However, the coefficient for the average number of children is negative, which could be caused by the correlation

Table 6-8 Results of Model with Detailed Location Variables and Linear Income variable (t-stat in the parenthesis)

	Slope Coefficient				Marginal Effects			
	Model 2		Model 3		Model 2		Model 3	
ONE	-4.9005	(-7.78)	-4.5509	(-6.06)	-0.9886	***	-0.9178	***
Inc	0.0022	(6.64)	0.0025	(7.65)	0.0005	***	0.0005	***
Child	-0.2576	(-1.71)			-0.0520	*		
Worker	0.2225	(2.34)			0.0449	**		
HHSize	0.0661	(0.41)			0.0133	'		
HH2			1.5915	(2.06)			0.3210	**
HH3			-1.3072	(-1.56)			-0.2636	'
HH4			-0.7008	(-1.38)			-0.1413	'
HH5			-0.5407	(-1.01)			-0.1090	'
HH6			-0.7864	(-1.73)			-0.1586	*
HH7			0.4941	(0.92)			0.0997	'
HH8			-0.1143	(-0.22)			-0.0231	'
Area2	2.3635	(3.77)	2.2550	(3.56)	0.4768	***	0.4548	***
Area3	1.1794	(1.87)	0.9623	(1.51)	0.2379	*	0.1941	'
Area4	1.7840	(3.05)	1.6578	(2.82)	0.3599	***	0.3343	***
Area5	1.3787	(2.48)	1.2387	(2.22)	0.2781	**	0.2498	**
Price	-0.0154	(-7.04)	-0.0130	(-5.42)	-0.0031	***	-0.0026	***
RunCst	0.0013	(0.78)	0.0016	(0.95)	0.0003	'	0.0003	'
Age	0.1360	(13.45)	0.1543	(11.57)	0.0274	***	0.0311	***
AgSq	-0.1494	(-13.07)	-0.1808	(-11.09)	-0.0301	***	-0.0365	***
LL	-49459		-49454					
Null LL	-59422		-59422					
Adj LRI	0.1674		0.1675					

Table 6-9 Income and Price Elasticity (Model 2 & 3 with linear income and price variable)

	Model 2			Model 3		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.60	-1.31	*0.14*	0.67	-1.10	*0.17*
Median	0.56	-1.04	*0.09*	0.61	-0.86	*0.11*
High	0.53	-0.88	*0.06*	0.58	-0.76	*0.08*

Note 1: the running cost elasticity is not statistically significant, hence in italic.
Note 2: Low income: 10 percentile of car owning household income, £251 per week;
Mid income: 50 percentile of car owning household income, £370 per week;
High income: 90 percentile of car owning household income, £481 per week.

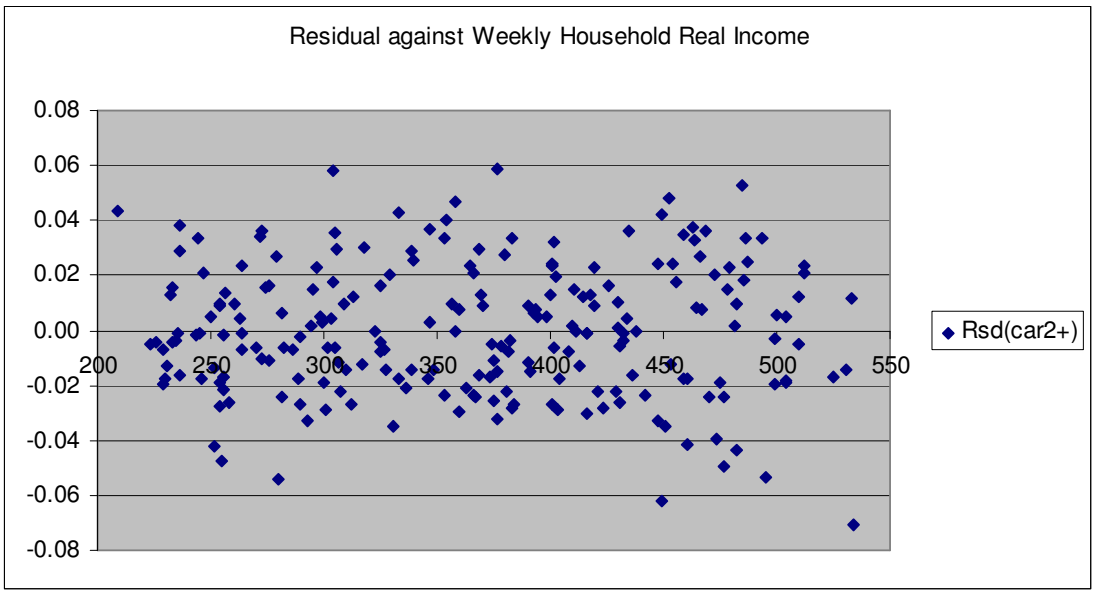
between the three household characteristics variables. When the eight-way categorization of household types is used, most of the household type variables are not significant. The coefficients for cohort age and square of cohort age (divided by 100 for scaling purpose) are highly significant, and their different signs indicate a peak of second (or more) car ownership in the cohort life cycle.

Regarding the impacts of income and costs on car ownership, it is clearer to discuss them in terms of elasticity. Table 6-9 reports the income and price elasticity for households with different income level.

The income elasticity of owning 2+ cars conditional on owning the first one is much higher than that of owning at least one car, especially for high income households. Based on Model 3, a 1% income rise would increase the conditional probability of owning two or more cars from 0.4384 to 0.4409 for households at 90 percentile income. For those at 10 percentile income, a 1% income rise would increase the conditional probability of second car ownership from 0.0588 to 0.0592. The purchase price elasticity for 2+|1+ cars is also higher than that for 1+ car, which is what we expect. While the ownership of at least one car might be a necessity for many households, owning the second or even more cars is certainly more of an option. The coefficients for running costs are not significant for both Model 2 and Model 3, so no reliable inference can be made on running cost elasticity.

Model 2 and 3 are in logit form, and we have also investigated the alternative functional form. Similar to the model of one plus car, the marginal effects from the probit model are almost identical to those from Model 2 and 3. Furthermore, it appears that heteroskedasticity is not an issue here, as illustrated by the residual plot of Figure 6-4.

Figure 6-4 Residual against Household Income (Model 3, Car 2+|1+)



The linear form of income variable implied that the impact of income change on utility is the same regardless of current income level. This assumption might not be appropriate, so we have examined alternative models with logarithm transformation of the income and price variables. These results are reported in Table 6-10.

Table 6-10 Model 2+|1+ with Log Income and Log Price Variables (t-stat in the parenthesis)

	Slope Coefficients				Marginal Effects			
	Model 5		Model 6		Model 5		Model 6	
ONE	-7.4219	(-3.02)	-8.8338	(-3.44)	-1.4969	***	-1.7812	***
LnInc	1.1154	(7.56)	1.2001	(8.53)	0.2250	***	0.2420	***
Child	-0.2864	(-1.88)			-0.0578	*		
Worker	0.1643	(1.68)			0.0331	*		
HHSize	0.0973	(0.60)			0.0196	'		
HH2			1.4199	(1.84)			0.2863	*
HH3			-1.3016	(-1.55)			-0.2624	'
HH4			-0.6160	(-1.20)			-0.1242	'
HH5			-0.7498	(-1.41)			-0.1512	'
HH6			-0.9917	(-2.19)			-0.2000	**
HH7			0.3316	(0.62)			0.0669	'
HH8			-0.3349	(-0.65)			-0.0675	'
Area2	2.5714	(4.14)	2.4443	(3.88)	0.5186	***	0.4929	***
Area3	1.4200	(2.26)	1.1951	(1.89)	0.2864	**	0.2410	*
Area4	1.9605	(3.35)	1.7983	(3.05)	0.3954	***	0.3626	***
Area5	1.5122	(2.73)	1.3856	(2.48)	0.3050	***	0.2794	**
LnPrice	-1.2825	(-5.46)	-0.9638	(-3.77)	-0.2587	***	-0.1943	***
LnRunCst	0.3035	(1.69)	0.3510	(1.88)	0.0612	*	0.0708	*
Age	0.1213	(11.20)	0.1436	(10.49)	0.0245	***	0.0290	***
AgSq	-0.1333	(-11.04)	-0.1697	(-10.24)	-0.0269	***	-0.0342	***
LL	-49463		-49455					
Null LL	-59422		-59422					
Adj LRI	0.1674		0.1674					

Comparing the results in Table 6-10 and Table 6-8, the marginal effects of the household characteristic, location and cohort age variables are very similar. The marginal effects of the income and price variables are not directly comparable. For high income households, the income elasticity derived from the log model is similar to that derived from the linear variable model; while for low income households, the log models lead to much higher income elasticity. All these results are very similar to the earlier discussion of the one plus car model. However, there is one distinctive difference for Model 2+|1+, i.e. the log models have the worse fit than the linear variables models. The log likelihood of Model 5 is reduced by 3.5 compared to Model 2, and that of Model 6 is lower by 1.3 when compared to Model 3. Nevertheless, the

difference in log likelihood is very small, and the log models can not be dismissed completely because they appear to have more realistic income elasticity profile.

Finally, the Fixed Effect Models and Random Effects Models have been examined. Unlike the model of one plus car, the introduction of cohort dummies in the Fixed Effect Model does not improve the model fit, and the log likelihood actually reduces in one case. The Random Effect Model does not lead to better model fit either, which is similar to the case of one plus car model. Further investigation of the fixed effect and random effect models will be presented in the next chapter.

6.5 Conclusion

In this chapter, we extend the pseudo panel method to the discrete choice case, which is the first study of the kind to our best knowledge. To understand why it has never been investigated in the literature, and why it could prove an effective technique for empirical work, it is necessary to discuss the pros and cons of nonlinear pseudo panel model method. Compared to the conventional cross sectional model, nonlinear pseudo panel has two main advantages: consideration of dynamics in modelling and effective tackling of aggregation bias problem. On the other hand, it has the disadvantages of reducing data variability and loss of information on individual decision makers. Compared to the more familiar linear counterpart, nonlinear pseudo panel method has the advantages of explicitly modelling and estimating the saturation level as well as consistency with the theory of utility maximization. However, it also suffers two limitations: biased estimation of the fixed effect models due to "incidental parameter problem" and the need for tailored code for estimating advanced models. On balance, nonlinear pseudo panel model is most suitable for forecasting purpose, and the case is less clear for analytical purpose.

We then go on to introduce a random utility model of pseudo panel. In a standard random utility model of cross sectional data, the utility function consists of a deterministic term and a random term. For pseudo panel model, the deterministic term can be further decomposed into three components: the sample mean observable utility, measurement error and decision-maker's utility deviation from cohort mean. The resulting random utility pseudo panel model has a similar probability function but with

different scale compared to the cross-sectional model. The pseudo panel RUM is then applied to car ownership modeling. The pros and cons of various modeling structures are evaluated and the hierarchical structure for handling multiple car ownership is subsequently chosen.

In this chapter, we also discuss the consistent estimation of the pseudo panel RUM. The fixed effect estimator is consistent only when the number of time periods is sufficiently large, while the random effect estimator requires the orthogonality assumptions that the unobserved heterogeneity are uncorrelated with the explanatory variables. While some authors have attempted to relax the orthogonality assumption of the random effect model by parameterizing the distribution of λ_i as a function of x_i, it is not always easy to justify the such parameterization if one want it to hold for all T_i. As a result, this approach is not adopted in the empirical work; instead, we rely on the large T consistency of the fixed effect estimator, justified on the ground that there are 19 periods for most cohorts in the pseudo panel dataset. The last section of this chapter reports the empirical results of car ownership model. Separate results are presented for models of one plus cars and those of two plus cars. Systematic specification search has been conducted to determine the models with the best fit, and the income, purchase price and running costs elasticity derived from these models are shown to be sensible and comparable to other studies.

Chapter 7 Dynamic Model and Model with Saturation

The importance of dynamics in transport analysis has been increasingly recognised in recent years and it is argued that travel behaviour should be treated as a process in continual flux, with history and a path through time (Goodwin, 1997). Dynamics in behaviour have been identified in the literature: people can anticipate the expected new situation or adapt their behaviour in the longer term. However, cross sectional data and models based on such data rely on assumptions that equilibrium is observable in the real world; when this is not the case, the cross sectional model would give a biased picture in people's choice behaviour (except under certain special conditions). This is the main reason that compels us to depart from the cross sectional model and adopt a pseudo panel approach.

On the other hand, we have argued that linear pseudo panel model is not sufficient when dealing with durable goods. There are two main reasons behind this argument, i.e. consistency with the micro economic theory of utility maximization and explicit consideration of saturation. While saturation can be approximated through certain linear transformation of variables (e.g. semi-log *X* models) in the linear model, with discrete choice model, it can be more accurately measured and its statistical significance can also be tested.

In this chapter, we discuss a dynamic random utility model of pseudo panel, its consistent estimation and applications to car ownership models. As discussed in previous chapter, the utility function for pseudo panel model mirrors that of individual decision-maker and the probability model of pseudo panel is similar to that of cross sectional models except for the different scale. As a result, it is more convenient to start the model development based on individual decision-maker, and extending it to pseudo panel would be a straightforward task. In Section One, we first present a general structural model encompassing three forms of choice dynamics; after detailed evaluation of each sub-model, the standard state dependence model (first order Markov model) is selected and we subsequently derive its pseudo panel version. Section Two

discusses the consistent estimation of the preferred structural model developed in Section One. Thanks to the close relationship between the random utility model of individuals (genuine panel of discrete data) and cohorts (pseudo panel of proportions data), we are able to draw on parametric and semi-parametric techniques developed for discrete data and propose appropriate estimation method for the proportions data.

Section Three reports the empirical finding of dynamic pseudo panel model of car ownership. In the fourth section, the focus of investigation is changed from dynamics to saturation, another motive for the use of non linear pseudo panel model. We present a "Dogit" type car ownership model, which incorporates saturation into the random utility model framework. Empirical results of models with saturation are reported subsequently. Section Five is a brief conclusion.

7.1 Dynamic Random Utility Model of Pseudo Panel

In the econometric literature, the importance of distinguishing true state dependence and spurious state dependence caused by unobserved heterogeneity has been widely recognised (e.g. Heckman, 1981a; Arellano and Honore, 2001; Greene, 2003; Carro, 2003). Receiving much less attention is one related topic, i.e. distinguishing different forms of *true* state dependence, which is the centre of our investigation here. Following a general to specific approach, we start from a general dynamic model, taking into account three prevalent forms of true state dependence. While the general model is a binary choice model based on discrete data, it will later be transformed to model of proportions data (pseudo panel) in section 7.1.4 and can also be extended to the case of multinomial choice.

According to the random utility model, the utility of choosing an option is the sum of a deterministic term and a random term unobservable to the researchers. If there is true state dependence, the choice made in one period will affect the utility in other periods. As a result, the deterministic utility component V_{it} would not only include the explanatory variables in the current periods, but also the past and future choices and past value of explanatory variables. For a general dynamic model, equation (1) describes the breakdown of the utility components for individual *i* choosing Option 1 in year *t*:

$$U_{it} = V_{it} + \varepsilon_{it} = \beta' x_{it} + \alpha \cdot y_{i,t-1} + \rho \cdot U_{i,t-1} + \sum_{j=1}^{\infty} \xi_{t+j} y_{i,t+j} + \varepsilon_{it} \qquad (1)$$

Then individual *i* in year *t* would choose Option 1 if and only if the utility of choosing Option 1 is greater than that of choosing the alternative option: $y_{it} = 1 \Leftrightarrow U_{it} > U'_{it}$. Normalising the utility of choosing the alternative option to zero ($U'_{it} = 0$), and the model can be formally expressed as:

$$y_{it} = \mathbf{1}\,(U_{it} > 0) \qquad (2)$$

Equation (1) and (2) describe the general dynamic random utility model for individual decision maker, which needs to be estimated using genuine panel data. As it will be straightforward to transform this model into a pseudo panel form at a later stage, it is more convenient to based our analysis on the individual decision-maker for now, and discuss the meaning, underlying economic theory and specific models represented by the various terms on the right hand side of equation (1).

The first term represents the effect of exogenous variables on current utility comparison. It is assumed that *X* is strictly exogenous, i.e. the past, current and future values of *X* are un-correlated with the error term ε_{it}. Meanwhile, the last term ε_{it} is the random utility component, representing the factors unobservable to researchers. When there is temporal correlation between ε in different periods, the dynamic effect would also be caused by the effect of unobserved heterogeneity or true series correlation. The lagged dependent variable might appear significant due to temporal correlation in the random term even there is no true state dependence. This is an important issue that has to be considered in model development.

While the second, third and fourth terms on the right hand side of equation (1) appear together in the general model, it is more common that they appear separately with the exogenous variables and the error term. The three corresponding types of dynamic models are: standard state dependence model, propensity dependence model and dynamic optimization model.

7.1.1 Standard State Dependence Model

In a standard state dependence model, states (choices) in the previous periods affect the choice in the current period. An example of standard state dependence in labour economics is the evidence of past work experience raising wage rates and thus raising the probability that a woman works in the future (Heckman 1981a). In the case of car ownership, state dependence arises because households that already own a car experience higher utility of car ownership in the current period because, for example, they develop a habit of relying on their car and they are more familiar with the road network and traffic conditions. When persistence is in the form of standard state dependence, its effects can be modelled by the inclusion of a lagged dependent variable, i.e. the second term on the right hand side of equation (1). When only the choice in the last period is included in the utility function, it is also called first order Markov model.

There are two potential extensions to the standard state dependence model with first order lagged effect. The first is Heckman (1981a), where he considered the effect of the entire past history of the process on current choice and the coefficients are assumed to be the functions of the current time period *t*, and the period in which the event occurred, *t-j*. The second is Beck et al. (2002), where it is shown that the state dependence model here is a special case of a "full" transition model, i.e. estimating two separate models conditional on the previous state (0 and 1 as in the current binary choice model). However, there appear to be no significant benefits in adopting either extension in the current study, so we refrain from making further investigation into these models.

7.1.2 Models of Propensity Dependence

When the lagged effect is represented by the utility difference in the previous period rather than the actual state in the previous period, the model become a propensity dependence model, also called Latent Markov model. It assumes that the prior propensity to select an option rather than prior choice itself determines the current probability that an option is chosen.

Models of propensity dependence have a closer resemblance to the time series model of exponential decay, capturing the idea that it takes time for a change in an independent variable to fully work its way through the system, while the most recent past receives the greatest weight. As shown in Beck (1991) and Beck et al. (2002),

applying the Koyck (1954) transformation[33] to an exponentially distributed lag model of (3):

$$U_{it} = \beta' x_{it} + \rho \cdot \beta' x_{i,t-1} + \rho^2 \cdot \beta' x_{i,t-2} + ... + \varepsilon_{it} + \rho \cdot \varepsilon_{i,t-1} + \rho^2 \cdot \varepsilon_{i,t-2} + ... \quad (3)$$

one would obtain the utility function of propensity dependence, with $U_{i,t-1}$ and various exogenous variables being the explanatory variables.

An alternative motivation for the exponential distributed lag model is the accumulation of information and formation of expectations (Greene, 2003). For example, in our car ownership model, households form expectations about their future income and it is assumed that the currently formed expectation is a weighted average of the expectations in the previous periods and the most recent observation. When the budget constrain is relaxed to account for future income growth (with borrowing), the conditional indirect utility function would include income expectation rather than actual income. Such model of expectation, after some manipulation, would lead to the exponential distributed lag model of (3).

Model of propensity dependence can be estimated using Bayesian method of Markov Chain Monte Carlo (MCMC), which should not be a big burden with modern computing. However, there is considerable difficulty in using the estimation results in forecasts, as the lagged utility is a latent variable and unobservable. Since the objective of the empirical study in this project is to develop car demand forecasts in Great Britain, the value of modelling propensity dependence seems limited in this context. On the other hand, the idea that income expectation rather than actual income should determine the indirect utility function of car ownership remains an interesting and credible hypothesis. The testing of competing hypothesis of state dependence and propensity dependence in household car ownership decision could be undertaken in future projects.

---

[33] The transformation is done by multiplying both sides of equation (3) by $(1 - \rho \cdot L)$ with *L* being the lag operator. Note that $\lim_{s \to \infty} \rho^s \cdot \beta' x_{i,t-s} = 0$, assuming a stationary condition of $\rho < 1$.

7.1.3 Models of Dynamic Optimisation

The model of dynamic optimisation is related to the concept of state dependence: if past choices affect current choice, then current choices affect future ones; consequently, a decision maker who is aware of this fact will take the future effects into considerations. A link from the past to present would imply a link from the present to the future (Train, 2003). In this case, it seems natural to include the effect of future choices on the current choice, and this effect is represented by the fourth term on the right hand side of equation (1).

Adding future choices to the utility function of (1) is a simplified example of the dynamic optimisation model. As a binary choice model, it does not consider the impacts of current choice on the attributes and availability of other alternatives in the future. When car ownership is examined as a standalone decision, it is difficult to see how owning a car or not in the future could influence the utility of car ownership now.

A well defined dynamic optimisation model assumes that decision makers choose the alternative in the current period that maximizes his expected utility over the current and future periods. For example, in Train (2003), when students decide whether or not to go to college, they consider not only the current period of college years, when the trade off is made between study and work, but also the post college years with different employment opportunities and even the post retirement years with different retirement options.

An interesting application of dynamic optimisation in the study of car ownership is Adda and Cooper (2000). They developed a dynamic transaction model of discrete choice to study the effect of government subsidies on durable goods market. By assuming a positive value to a car in any age and no used car market, the choice set is reduced to two alternatives in each period: keeping the existing car of age i; or replacing it with a new one and receiving a government subsidy. For each option, the total utility is the sum of three components:

1) current utility flow;
2) probability-weighted (discounted) expected utility of using the car (of age i or age 2 depending on the current choice) in the next period;

3) probability-weighted (discounted) expected utility of scrapping the car owned in the current period and replacing it with a new car in the next period.

The optimisation problem assumes no borrowing or lending and a stochastic process for income, prices and aggregated taste shocks. Based on aggregate data, the authors use observations on sales as well as moments of the cross-sectional distribution to identify the parameters for the dynamic programming problem.

Model of dynamic optimisation offers an alternative approach in modelling car ownership dynamics. This approach is not adopted in the current study, because it requires a fundamental shift of the underlying car ownership model from holding model (as examined in the current study) to transactions model. Nevertheless, it could be a fruitful area for future research.

Finally, as noted by Train (2003, p174), "a model of rational decision making over time does not necessarily represent behaviour more accurately than a model of myopic behaviour, where the decision maker ignores future consequences". To turn a complex dynamic optimisation problem into a tractable form, one usually has to impose certain restrictive assumptions, which would reduce the appeal of such models in empirical work.

7.1.4 Transforming the Reduced Model for Repeated Cross Sections

In the previous sections, we have discussed a general model that encompasses three specific types of structural state dependence model: standard state dependence model, propensity dependence model and dynamic optimisation model. Based on the pros and cons of each model and their relevance to the car ownership forecasting, it seems appropriate to select the standard state dependence model for the current empirical work. In this model, household i in year t choose to own a car when such ownership yields a positive utility U_{it}, if the utility of not owning car is normalised to zero. Formally, the car ownership model can be expressed as:

$$y_{it} = \mathbf{1}\,(U_{it} > 0) \qquad (4)$$

where the utility function in our reduced model is:

$$U_{it} = V_{it} + \varepsilon_{it} = \beta' x_{it} + \alpha \cdot y_{i,t-1} + \varepsilon_{it} \qquad (4a)$$

However, model (4) is still based on genuine panel data. For repeated cross sectional data, different individuals are sampled in different years, so notation similar to that of the linear models in Chapter 4 and 5 has to be used. Adding the time dimension to the person identifier, *i* becomes *i*(*t*) and equation (4a) can be written more precisely as:

$$U_{i(t),t} = V_{i(t),t} + \varepsilon_{i(t),t} = \beta' x_{i(t),t} + \alpha \cdot y_{i(t),t-1} + \varepsilon_{i(t),t} \qquad (4b)$$

For repeated cross section data, household's choice in the previous period, $y_{i(t),t\text{-}1}$, is unobservable; instead, we only have information of $y_{i(t\text{-}1),t\text{-}1}$. In order to investigate the choice dynamics, repeated cross sectional data have to be aggregated into pseudo panel[34]. Assuming no birth or death in the total population and defining cohorts based on time-invariant variables, the cohort population remains fixed over time. As a result, we can write the deterministic part of the utility function in (4b) as true cohort population mean plus deviation from such mean ($\theta_{i(t),t}$) for individual *i* in year *t*:

$$\beta' x_{i(t),t} + \alpha \cdot y_{i(t),t-1} = \frac{1}{N_c} \sum_{i=1}^{N_c} (\beta' x_{it} + \alpha \cdot y_{i,t-1}) + \theta_{i(t),t}, \forall i \in c \qquad (5)$$

However, the true cohort population mean of the deterministic utility components are unobservable; instead, we only have cohort sample mean calculated from two consecutive years. Note the total measurement errors in these two periods as $(\eta_{ct} + \eta_{c,t-1})$, we have:

$$\frac{1}{N_c} \sum_{i=1}^{N_c} (\beta' x_{it} + \alpha \cdot y_{i,t-1}) = \frac{1}{n_{ct}} \sum_{i(t)=1}^{n_{ct}} (\beta' x_{i(t),t}) + \frac{1}{n_{c,t-1}} \sum_{i(t-1)=1}^{n_{c,t-1}} (\alpha \cdot y_{i(t-1),t-1}) + \eta_{ct} + \eta_{c,t-1} \qquad (6)$$

Finally, we turn our attention to the error term $\varepsilon_{i(t),t}$. Since it is important to distinguish true state dependence and the so-called "spurious state dependence"[35], where the

[34] Another possibility is to use the first order Markov models proposed in Moffitt (1993), in particular the linear probability model for hazards. However, the data requirement for such model is very high, as it requires the previous values of the explanatory variables, or at minimum, the accurate backcast of such variables.

[35] As pointed out by Heckman (1981a), the lagged dependent variable in the dynamic model might appear significant even if there is no true state dependence. In another word, inter-temporal correlation of the error term has to be accounted for before true state dependence can be revealed.

dynamic effect is caused by unobserved heterogeneity, we assume a "components of variance" structure of the error term[36]:

$$\varepsilon_{i(t),t} = \lambda_c + \varepsilon'_{i(t),t} \qquad (7)$$

where λ_c is the (time-invariant) unobserved heterogeneity (cohort fixed or random effect) and is assumed to distributed independent of $\varepsilon'_{i(t),t}$;

$\varepsilon'_{i(t),t}$ captures the randomness besides heterogeneity, which is assumed to be independently identically distributed with mean zero and variance σ^2.

Substituting (5), (6) and (7) into equation (4b), and noting the cohort sample mean of the deterministic utility component as $\overline{V}_{ct}$, we have:

$$U_{i(t),t} = \overline{V}_{ct} + \eta_{ct} + \eta_{c,t-1} + \theta_{i(t),t} + \lambda_c + \varepsilon'_{i(t),t} \qquad (8)$$

where $\overline{V}_{ct} = \frac{1}{n_{ct}} \sum_{i(t)=1}^{n_{ct}} (\beta' x_{i(t),t}) + \frac{1}{n_{c,t-1}} \sum_{i(t-1)=1}^{n_{c,t-1}} (\alpha \cdot y_{i(t-1),t-1})$.

Similar to the static model in the previous chapter, two simplifications have been applied to the utility function (8). This first is about the measurement errors. Under the asymptotic of $n_{ct} \to \infty, \forall t$, the measurement errors converge in probability to zero:

$$p\lim(\eta_{ct} + \eta_{c,t-1}) = 0 \qquad (9)$$

In another word, when the cohort sample size is sufficiently large, which is the case for the current project, the measurement errors can be ignored.

The second is the aggregation of two sources of randomness into a composite error term. Residual error $\varepsilon'_{i(t),t}$ and deviation from the true cohort mean (deterministic) utility $\theta_{i(t),t}$ are aggregated, as they are empirically in-distinguishable in the pseudo panel setting:

$$e_{i(t),t} = \theta_{i(t),t} + \varepsilon'_{i(t),t} \qquad (10)$$

---

[36] In theory, there is another source of inter-temporal correlation of errors: series correlation of the residual error term ε'. This is ignored in the current study for the reason suggested by Beck et al. (2002), i.e. series correlation might not be important after accounting for dynamics.

Substituting (9) and (10) into (8), we obtain the utility function of a discrete choice pseudo panel model with state dependence (equation 11). Note that the deterministic utility component and unobserved cohort heterogeneity are the same for all decision makers in cohort *c*, and only the composite error term $e_{i(t),t}$ is different across decision makers. The random term $e_{i(t),t}$ causes the actual (observed) choice behavior to vary between individuals in the same cohort.

$$U_{i(t),t} = \overline{V}_{ct} + \lambda_c + e_{i(t),t} \qquad (11)$$

Conditional on unobserved heterogeneity λ_c, one can estimate model (11) in the form of probit or logit, depending on the assumption on the distribution of $e_{i(t),t}$. In the following section, we will discuss the consistent estimation of the dynamic discrete choice model based on utility function of (11).

7.2 Consistent Estimation of Dynamic Model

In the previous section, we investigate a structural model of discrete choice with different forms of true state dependence. The standard state dependence model has been selected as the preferred dynamic model of car ownership, which probably is the simplest among all three sub-model types. However, the consistent estimation of the standard state dependence model is not a trivial issue, even for genuine panel data. To obtain unbiased estimation of the structural parameters, it is important to appropriately account for the unobserved heterogeneity (cause of "spurious state dependence"). For genuine panel, there is a growing literature on the fixed effect, random effect and semi-parametric estimators of the standard state dependence model (first order Markov model). Because the discrete choice models of pseudo panel and genuine panel share similar utility and probability function, it would be beneficial to first review these estimators. Based on the findings of such literature review, we will propose the parametric methods to be used for the current study and discuss their application to pseudo panel.

The difficulty of consistent estimation arises from the inclusion of lagged dependent variable in the explanatory variables. This has different consequences on the fixed effect models and the random effect models. For fixed effect model, the presence of

lagged dependent variable increases the bias caused by the "incidental parameter problem". In the Monte Carlo study of Heckman (1981b), the bias of fixed effect estimator for $T = 8$ varies between 12% and 40% depending on the assumption of variance and "true" parameter value. Such bias is significantly larger than the bias of up to 10% for static fixed effect model cited in the same study.

For random effect model, the inclusion of lagged dependent variable causes a difficult "initial conditions problem". If the first sample observation is a state during a process, then it would depend on the dependent variable before the sampling period, although this is usually tackled by dropping the first observation in empirical work. The real difficulty lies in the relationship between the first lagged dependent variables and the unobserved heterogeneity[37], which depend on the parameters of the model as well as the distribution of the explanatory variables in periods prior to the start of the sample (Arellano and Honore, 2001). In many empirical studies, the initial condition problem is ignored by assuming the first lagged dependent variable is strictly exogenous. This assumption is justifiable only when one can reasonably assume the process is observed from the start. For example, in labour economics, if the sample period starts from people leaving secondary school, then the first observation of labour participation can be assumed to be strictly exogenous and no initial condition problem would arise.

7.2.1 Literature Review on Genuine Panel Model

In this sub-section, we review the dynamic discrete choice models with fixed effects and random effects proposed in the literature. While the emphasis is on parametric methods due to a higher degree of relevance to the current theme of demand forecasts, some important semi-parametric methods will also be reviewed for reference purpose.

7.2.1.1 Fixed Effect Models and Incidental Parameter Problem

In the last chapter, we discussed the incidental parameter problem that causes bias in the maximum likelihood estimation of the fixed effect estimators when the number of time period is fixed. The conditional maximum likelihood estimator, while achieving consistency for large C, does not allow the calculation of marginal effect, which

[37] Unlike the linear dynamic panel data model, the unobserved heterogeneity in discrete choice model can not be eliminated by taking first difference.

severely limits its usefulness in empirical work. Given the attractiveness of the fixed effect estimator (the distribution of the unobserved heterogeneity is left un-specified and no orthogonality assumption between the explanatory variables and error terms), another class of the estimator was proposed in the literature, which seeks to reduce the order of bias rather than achieving fixed T consistency. This is the modified concentrated likelihood estimator, initialled proposed in Arellano (2003) and extended by Carro (2003) to the dynamic model.

The modified concentrated likelihood estimator is implemented in three steps: first is the reparametrization of the original parameters (γ, λ_i) in the log-likelihood function $l_i(\gamma, \lambda_i)$ to (γ, ς_i), so that $l_i(\gamma, \lambda_i(\gamma, \varsigma_i)) = l_i^*(\gamma, \varsigma_i)$,

where $\gamma = (\alpha, \beta')'$ and $E\left(\frac{\partial^2 l_i^*(\gamma_0, \varsigma_{0i})}{\partial \gamma \partial \varsigma_i}\right) = 0$ is satisfied[38]. The condition that the *expected* cross derivative is zero implies information orthogonality between the structural parameters γ and the nuisance parameters ς_i. Information orthogonality is a weaker condition than standard form of orthogonality, which requires the cross derivative (rather than its expected value) to be zero and can not be achieved for discrete choice model.

The second step is to modify the concentrated likelihood, because the maximum likelihood estimator of the γ does not change with the reparametrization and still have bias of order $O(T^{-1})$. The modified concentrated likelihood follows Cox and Reid (1987) and is expressed as:

$$l_{Mi}(\gamma) = l_i^*(\gamma, \hat{\varsigma}_i(\gamma)) - \frac{1}{2}\log[-d_{\varsigma\varsigma}^*(\gamma, \hat{\varsigma}_i(\gamma))] \qquad (12)$$

where $\hat{\varsigma}_i(\gamma)$ is the maximum likelihood estimator of ς_i given γ, and the modification term is a log function of $d_{\varsigma\varsigma}^*(\gamma, \hat{\varsigma}_i(\gamma)) = \partial^2 l_i^* / \partial \varsigma_i^2$. The modification term is introduced to penalize values of γ for which the information about the effects is relatively large.

[38] Subscript *0* denotes the true parameter values. Also, it should be noted that in the original paper of Arellano (2003) and Carro (2003), the unobserved heterogeneity is noted as η_i while the reparametrization is noted as λ_i. The notation is changed to be consistent with the rest of the thesis.

The third step involves re-writing (12) in terms of original parameterization. The modified concentrated likelihood of (12) is not directly usable, as it is based on reparametrization form of $l_i^*(\gamma, \varsigma_i)$. Expressing the term $d_{\varsigma\varsigma}^*(\gamma, \hat{\varsigma}_i(\gamma))$ as the product of Fisher information in the (γ, λ_i) parametrization and the square of the Jacobian of the transformation from (γ, λ_i) to (γ, ς_i), (12) becomes:

$$l_{Mi}(\gamma) = l_i(\gamma, \hat{\lambda}_i(\gamma)) - \frac{1}{2}\log[-d_{\lambda\lambda}(\gamma, \hat{\lambda}_i(\gamma))] + \log\left(\frac{\partial \varsigma_i}{\partial \lambda_i}\bigg|_{\lambda_i = \hat{\lambda}_i(\gamma)}\right) \qquad (13)$$

The modified concentrated likelihood function of (13) can be further simplified for binary logit and probit model and estimated using maximum likelihood estimation. The modified concentrated likelihood estimator, although not fixed *T* consistent, reduces the bias of the estimated parameters from $O(T^{-1})$ to $O(T^{-2})$. Monte Carlo experiments in Carro (2003) shows that the bias is sufficiently small for logit and probit model with lagged dependent variable and strictly exogenous variables. Furthermore, the fixed effect λ_i is calculated during the estimation process, which enables the calculation of the marginal effect of the explanatory variables. This makes the modified concentrated likelihood estimator more attractive compared to other fixed *T* consistent estimators.

7.2.1.2 Random Effect Model and Initial Condition Problem

After reviewing the parametric model proposed in the literature to tackle the incidental parameter problem of the fixed effect model, we turn our attention to the initial condition problem of the random effect model. As mentioned before, in a standard first order Markov model, the distribution of the process $(y_{i1}, ..., y_{iT})$ can be specified conditional on the strictly exogenous variables, the individual specific effect (unobserved heterogeneity) and the initial condition y_{i0}, which can also be the first lagged dependent variable. However, the model does not specify the distribution of y_{i0} conditional on the individual effects and the strictly exogenous variables. There are essentially two parametric approaches in tackling the initial condition problem. The first approach is proposed in Heckman (1981b), which involves specifying a separate model for y_{i0} given the unobserved effects and the strictly exogenous variables. More specifically, if one specifies $f(y_{i0} \mid x, \lambda_i)$, then

$$f(y_{i0}, y_{i1},..., y_{iT} \mid x, \lambda_i) = f(y_{i1},..., y_{iT} \mid y_{i0}, x, \lambda_i) \cdot f(y_{i0} \mid x, \lambda_i) \quad (14)$$

If the density of $f(\lambda_i \mid x)$ is specified, one can then integrate (14) with respect to this density to obtain $f(y_{i0}, y_{i1},..., y_{iT} \mid x)$. This model would in theory enable one to achieve consistent estimation by maximum likelihood, although the computation costs involved can be significant.

The second approach is proposed in Wooldridge (2005), which specify the distribution of the unobserved heterogeneity as $f(\lambda_i \mid y_{i0}, x)$, conditional on the strictly exogenous variables and the first observation y_{i0}. With prior assumptions made on the density of $(y_{i1},..., y_{iT})$ conditional on (y_{i0}, x, λ_i), assumptions on $f(\lambda_i \mid y_{i0}, x)$ can lead to the identification of the density of the entire process $(y_{i0}, y_{i1},..., y_{iT})$ conditional on the strictly exogenous variables *x*.

As shown in Wooldridge (2005), specifying $f(\lambda_i \mid y_{i0}, x)$ can lead to very tractable functional forms for some common nonlinear models, and the author argues that such specification is no worse than specifying separate models of initial conditions, which themselves can only be approximate. The approach suggested by Wooldridge is simple to implement and computationally attractive. It has been applied in dynamic car ownership model by Leth-Petersen and Bjorner (2005) using Danish household panel data. As a result, it would thus seem natural to use this approach in the current study.

However, such conclusion is premature for two reasons. Firstly, for unbalanced panel data, one has to specify a different conditional distribution of λ_i for each configuration of the missing data; alternatively, one can use the balanced sub panel if sample selection is exogenous conditional on (y_{i0}, x). The obvious drawback of the latter treatment is the danger of discarding a lot of useful information. The treatment of unbalanced panel is highly relevant in the current study as the pseudo panel dataset is unbalanced. Secondly, it is not realistic to assume a simple distribution of $f(\lambda_i \mid y_{i0}, x)$, as pointed out by Honore (2002). If the first observation depends on the random effect λ_i, then the distribution of λ_i conditional on (y_{i0}, x_{it}) will depend on the values of x_{it} before the start of the sample. The distribution of the random effect conditional on the

strictly exogenous variables and initial conditions will therefore depend on the time series properties of x_{it} in some very complicated way. This difficulty is confronted by Arellano and Carrasco (2003) in their semi-parametric model, which will be discussed below.

7.2.1.3 Semi-Parametric Model

Semi-parametric methods have the advantage of allowing estimation of parameters without specifying a distribution (conditional or unconditional) of the unobserved effects. Avoiding certain strong parametric assumptions, some semi-parametric models achieve consistent estimation of structural parameters based on minimum or no assumptions on nuisance parameters. Chamberlain (1985) considered a logit model with the lagged dependent variable as the only explanatory variables. Considering only the first four observations, the probability of not selecting a state in the second period ($y_{i1} = 0$) conditional on state-switching between the second and third period ($y_{i1} + y_{i2} = 1$) would depend on the state of the initial and fourth period ($y_{i0} - y_{i3}$) but not on the nuisance parameters. The result can be extended to *T* periods and be used to make inference on the structure parameters. Similar models were proposed in Jones and Landwehr (1988) and Magnac (1997), with the latter being a multinomial logit model.

Honore and Kyriadizou (2000) consider a model with lagged dependent variable as well as strictly exogenous variables. Considering only four periods, the authors showed that:

$$P(y_{i1} = d_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3 \mid x_i, \lambda_i, y_{i1} + y_{i2} = 1, x_{i2} = x_{i3})$$

$$= \frac{1}{1 + \exp((x_{i1} - x_{i2})\beta + \alpha(d_0 - d_3))} \qquad (15)$$

which does not depend on nuisance parameter λ_i . With continuous variable, the condition of $x_{i2} = x_{i3}$ in (15) is usually not satisfied. As a result, the author proposed a kernel estimator to average over observations close to the value. The major limitation of this approach is that as it uses only observations in a neighborhood of $x_{i2} = x_{i3}$, it is necessary to assume the x_{it} process satisfies $P(x_{i2} = x_{i3}) > 0$, which rules out time dummies or other variables that always increase for each cross-sectional unit.

In contrast to the fixed effect approach of Chamberlain (1985) and Honore and Kyriadizou (2000), the approach used by Arellano and Carrasco (2003) lies between fixed effect and random effect. The distribution of a composite error, $\varepsilon_{it} + \lambda_i$ conditional on all observables up to year t[39] is assumed to be homoskedastic normal. However, the unobserved heterogeneity is allowed to be correlated with the predetermined variables through the unspecified (non-parametric) conditional mean, thus avoiding the orthogonality conditions required by the standard random effect model. A Generalized Method of Moment (GMM) estimator was proposed for this model, which was shown to be consistent and asymptotically normal for fixed *T* and large *N*. One weakness of this approach, as recognized by the author, is that it matters to the distributional assumption if one starts observing the individuals one period earlier or later.

Besides the limitation of each specific model, there has been more general criticism of the semi-parametric models. As pointed out by Greene (2001a), the structural parameters in these semi-parametric models are essentially uninformative, as "they are not slopes of conditional means so they do not necessarily help in understanding behaviour". Also, the conditional means are unspecified, which renders these structural parameters in effect useless for prediction. This view is shared by Wooldridge (2005), who states that as the partial effects on the response probability or conditional mean are not identified, the economic importance of covariates, or even the amount of state dependence, cannot be determined from semi parametric approaches[40].

7.2.2 Estimation Methods Proposed for the Current Study

The literature review shows that the consistent estimation of discrete choice model with state dependence is a complex issue and different approaches have their advantages and limitations. For fixed effect models, it is not necessary to specify any distribution of the unobserved heterogeneity conditional on the explanatory variables, thus avoiding the strong orthogonality assumptions. For this reason, the fixed effect model should be the

[39] This condition determines the explanatory variables as predetermined (with feedback effect from the errors in the previous periods) rather than strictly exogenous.
[40] Another reason preventing us adopting semi-parametric approach is that it seems all but impossible to extend these methods from genuine panel data to pseudo panel.

preferred choice. However, treating the unobserved effects as parameters and estimating the model using maximum likelihood method produce biased estimation when the number of time period T is small. In the case of dynamic model, the bias appears to be larger with the presence of lagged dependent variables. Furthermore, the marginal effects in discrete choice model depend on the nuisance parameters, so eliminating these parameters, as in the case of certain semi-parametric models, does not seem to be the solution here.

For this reason, the approach of Arellano (2003) and Carro (2003), which aims at reducing the order of bias rather than achieving fixed T consistency, appears to be the most attractive. Because their approach is based on modifying the concentrated likelihood, there is no reason why this approach can not be extended to the pseudo panel model by adapting the log likelihood function to one based on proportions data rather than discrete data. However, it has proved very complicated in implementation and we have to leave it to future research. In the empirical work of the current project, we would rely on a rather optimistic result of the fixed effect estimator, i.e. consistency under the asymptotic of large T[41]. This action is partially justifiable on the grounds that we have relatively long sample periods of up to 19 years.

Regarding the random effect model, the main criticism is the strong orthogonality assumption between the unobserved effects and the explanatory variables. Furthermore, the relationship between the first lagged dependent variable and the unobserved heterogeneity is usually undefined, which leads to the so-called initial condition problem. Two solutions to the initial condition problem have been proposed in the literature, i.e. assumption on a separate distribution of the initial conditions in Heckman (1981b), or assumption on the distribution of the unobserved effects conditional on the initial conditions in Wooldridge (2005). However, neither approach is the definitive answer, as both have the real danger of modelling the distributions that are inconsistent with the data generating process.

[41] While consistency is established for genuine panel, it is likely to carry over to pseudo panel when measurement errors can be ignored.

In the current study, we propose a slightly different approach, i.e. random parameter models. Unlike the random effect model, where a constant term is assumed to capture the unobserved heterogeneity, such unobserved effects are captured by a randomly distributed parameter vector in a random parameter model. Adding a subscript i $(i = 1,...,n)$ to recognize heterogeneity in the parameter vector $\gamma = (\alpha, \beta')'$, Model (4) can be re-written as:

$$y_{it} = \mathbf{1}\left((y_{i,t-1}, x_{it}')\gamma_i + \varepsilon_{it} > 0\right) \qquad (16)$$

where, $\gamma_i = \gamma + \Gamma\nu_i$ is a vector of random parameter with mean γ and variance $\Gamma\Gamma'$;

Γ is a diagonal parameter matrix[42];

ν_i is a random vector with zero vector mean and covariance matrix $\mathbf{I}$.

Any fixed (non-random) parameters in γ_i can be specified by constraining the corresponding rows in Γ to be zero, and it is easy to see the random effect model is a special case of (16), where only the constant term is random. Recall the problem of random effect specification for dynamic discrete choice model. The distribution of the random effect is specified conditional on the explanatory variables, and it requires orthogonality assumption between the unobserved effect and the explanatory variables. With the presence of lagged dependent variable, the orthogonality assumption would usually be violated except for some special circumstances (e.g. the initial conditions are strictly exogenous and the unobserved effect is conditional on such initial conditions). However, in a random parameter model, the orthogonality condition becomes moot, as the individual specific heterogeneity is embodied in the marginal responses (parameters) of the model[43] (Greene, 2001a).

If one assumes the error term ε_{it} in (16) follows an iid logistic distribution, the underlying probability model has a logit form and (16) becomes a special (binary choice) version of mixed logit model. As shown in McFadden and Train (2000), mixed logit is a highly flexible model that can be used to approximate any random utility model. It alleviates the three limitation of standard logit by allowing for random taste

---

[42] The assumption of diagonal Γ makes our model a specific version of the general model in Greene (2004).

[43] Random parameter model also partially incorporate the idea of Heckman (1981b). Assuming the element corresponding to $y_{i,t-1}$ in the random vector ν_i is distributed as $N(0,1)$, Model (16) would encompass a standard random effect model with the initial condition assumed to be $y_{i0} \sim N(0,1)$.

variation, unrestricted substitution patterns and correlation in unobserved factors over time (Train, 2003). In recent years, mixed logit model has been applied to many empirical studies based on panel data (see, for example, Revelt and Train, 1998; Bhat, 2000; Mohammadian and Miller, 2003; Leth-Petersen and Bjorner, 2005, with the last two study being dynamic mixed logit model of car ownership) as well as cross sectional data (e.g. Bhat, 1998; Brownstone and Train, 1999; Hess and Polak, 2005; Hess et al. 2006; various mixed logic models of car ownership including Brownstone et al., 2000; Whelan, 2003; Golounov el al., 2004). One contribution of this study is to extend the application of mixed logit model to pseudo panel data.

For panel data model, dependence of the T_i observations for individual i results from the common random vector ν_i. Conditional on the unobserved heterogeneity, the observations are independent. In another word, conditional on ν_i (or alternatively on γ_i) the probability of a person makes a sequence of choices over T periods is the products of logit formulas:

$$p_{it}(\gamma_i) = \prod_{t=1}^{T} \frac{\exp((y_{i,t-1}, x_{it}')\gamma_i)}{1+\exp((y_{i,t-1}, x_{it}')\gamma_i)} \qquad (17)$$

since ε_{it}'s are independent over time. The unconditional probability is the integral of this product over all values of γ_i:

$$P_{it} = \int_{\gamma_i} p_{it}(\gamma_i) f(\gamma_i \mid \gamma, \Gamma) d\gamma_i \qquad (18)$$

Equation (17) and (18) illustrate the flexibility of mixed logit model in handling dynamic discrete choice model, as lagged dependent variables have been readily accommodated in the utility function in a given period to represent lagged response behaviour. Conditional on γ_i, the only remaining random terms in the mixed logit are the ε_{it}'s, which are independent over time, so the lagged dependent variable in the utility function is uncorrelated with these remaining error terms for period t (Train, 2003). In another word, the lagged dependent variable can be added to the mixed logit model without having to change the estimation procedure.

If we only have repeated cross sectional data rather than genuine panel data, transformation similar to those described in section 7.1.4 of this chapter has to be applied. After aggregating individual observation into cohorts, the utility of choosing

Option 1 by a particular individual in cohort *c* in year *t* can be expressed as the sample average (deterministic) utility $\overline{V}_{ct}$ plus various error terms. When the sample size is sufficiently large for each cohort, the measurement errors η_{ct} can be ignored. Deviation of individual deterministic utility from cohort mean, $\theta_{i(t),t} = V_{i(t),t} - \overline{V}_{ct}$, can be merged with the residual error term $\varepsilon'_{i(t),t}$. The only difference from Section 7.1.4 is the treatment of unobserved heterogeneity, λ_c, which will be absorbed into the random parameter vector γ_c in the mixed logit model (Note $\gamma_c = (\alpha_c, \beta'_c)' = \gamma + \Gamma \nu_c$ and assume that x_{ct} contains a constant term). Conditional on random vector γ_c, the sample average deterministic utility for cohort *c* will become the argument to the exponential function in (17), and the probability of any individual within cohort *c* choosing Option 1 over *T* period can be expressed as a product of logit formulas $\Lambda_{ct}^{\gamma_c}$ (raised to the power of $n_{ct} \cdot r_{ct}$):

$$p_{ct}(\gamma_c) = \prod_{t=1}^{T} [\frac{\exp(\overline{V}_{ct})}{1+\exp(\overline{V}_{ct})}]^{n_{ct} \cdot r_{ct}} = \prod_{t=1}^{T} (\Lambda_{ct}^{\gamma_c})^{n_{ct} \cdot r_{ct}} \tag{19}$$

where $\overline{V}_{ct} = \frac{1}{n_{ct}} \sum_{i(t)=1}^{n_{ct}} (\beta'_c x_{i(t),t}) + \frac{1}{n_{c,t-1}} \sum_{i(t-1)=1}^{n_{c,t-1}} (\alpha_c \cdot y_{i(t-1),t-1})$.

Similar to the case of genuine panel, the unconditional probability P_{ct} is the integral of (19) over all values of γ_c. The mixed logit model of (19) can be estimated most conveniently using the method of Maximum Simulated Likelihood (MMSL), which is based on the simulated logit probability within the right hand side of equation (19):

$$\tilde{\Lambda}_{ct} = \frac{1}{D} \sum_{d=1}^{D} \Lambda_{ct}(\gamma^d) \tag{20}$$

where γ^d is a value obtained from the *d*th draw of $f(\gamma_c \mid \gamma, \Gamma)$ and *D* is the number of draws. By construction, $\tilde{\Lambda}_{ct}$ is an unbiased estimate of Λ_{ct}. Inserting the simulated probability (20) into the log likelihood function of binary choice model based on proportions data, it gives a simulated log likelihood function:

$$SLL = \sum_{c=1}^{C} \sum_{t=1}^{T} n_{ct} [r_{ct} \ln(\tilde{\Lambda}_{ct}) + (1 - r_{ct}) \ln(1 - \tilde{\Lambda}_{ct})] \tag{21}$$

where n_{ct} is the sample size for cohort *c* and r_{ct} is the proportion of household in cohort *c* choosing Option 1 in year *t*.

For empirical application, the random parameter model for pseudo panel based on equation (19) to (21) has been implemented in Gauss (Aptech Systems, 1996). The code was adapted from a mixed logit program of Revelt and Train (1998), and various checks have been applied to ensure the correct implementation of the model. Appendix 2 is simplified version of the Gauss code we have used.

7.3 Empirical Results of Dynamic Car Ownership Model

Similar to the previous chapter, two separate pseudo panel datasets have been used to estimate the model of one plus car and that of two plus cars. For the former, the dataset was compiled using the entire sample of Family Expenditure Survey covering the period of 1982 to 2000. We exclude observations that are derived based on less then 100 households, resulting in a pseudo panel with 254 observations with 16 cohorts. The first observation of each cohort has to be dropped for dynamic model, so the number of pseudo panel observations is reduced to 238.

For the model of two plus cars, the dataset was constructed using a sub-sample of car owning households in the Family Expenditure Survey for the same period. After excluding observations based on small number of households, the pseudo panel dataset has 220 observations from 14 cohorts; for dynamic model, the sample size is further reduced to 206 after dropping the first observation of each cohort.

We carry out systematic specification search to determine the model with best fit. For clarity, results are reported separately for models of one plus car and two plus cars.

7.3.1 Dynamic Model of One plus Car

The explanatory variables in the utility function include the lagged dependent variable and other exogenous variables, including income and other household characteristics (household size, number of person in work, number of children), location of household as proxy for accessibility, costs of car ownership and use, and finally the second polynomial of average age of household head in the cohort.

In the initial tests, unobserved heterogeneity is not modelled and the models estimated are pooled logit or probit. We compared models with different representation of the

household characteristics, using either average household demographic statistics (household size etc.) or split of eight household types. We also compared models with different representation of the household locations, using either the full five location categories or the three compressed location categories. Regarding household income and costs of car ownership and use, variables of both linear form and logarithm form have been tested. Finally, different functional forms of logit and probit have been investigated. Table 7-1 summarises the various models mentioned above and compares their degree of fit based on log likelihood.

Table 7-1 Summary of Initial specification search

	Functional Form	**HH Characteristic Variables**	**Location Variables**	**Income/Cost Variables**	**No. of Variables**	**Log Likelihood**
Model 1	Logit	Ave No. of people*	Compressed	Linear	12	-65911
Model 2	Logit	Household types	Full	Linear	18	-65879
Model 3	Logit	Household types	Compressed	Linear	16	-65879
Model 4	Logit	Ave No. of people*	Full	Linear	14	-65910
Model 5	Logit	Ave No. of people*	Compressed	Log	12	-65901
Model 6	Logit	Household types	Compressed	Log	16	-65877
Model 7	Probit	Ave No. of people*	Compressed	Linear	12	*-65914*
Model 8	Probit	Household types	Compressed	Linear	16	*-65881*

(* Average number of people within the household, in work and under the age of 16)

The null log likelihood of the logit model is -81240, so the Adjusted Likelihood Ratio Index (also called Rho bar square) of Model 1 to 6 vary between 0.1885 and 0.1889. Using proportions of eight household types (Model 3) instead of average number of people (Model 1) as explanatory variables, the model loses 4 degree of freedom but has the log likelihood increased by 32 and the Adjusted Likelihood Ratio Index increased by 0.0004. On the other hand, there is almost no change of log likelihood when the five household location variables (Model 2) are compressed into three (Model 3), indicating no loss of fit. From these analysis, it appears that household characteristics are better represented by the eight-way categorization of household type, while locations are better represented by three area types, i.e. metropolitan areas, least populated rural areas and others.

Model 5 transforms the income and cost variables in Model 1 to the logarithm form, and the log likelihood is increased by 10. Similarly, Model 6 is the logarithm version of Model 3, while the log likelihood is increased by 2. As the degree of freedom is the

same between each pair of linear and logarithm models, the higher log likelihood suggests that models with log transformed income and costs variables are preferred.

Finally, Model 7 and 8 are the probit version of Model 1 and Model 3 respectively. Because the null log likelihood for the probit model is slightly different, the log likelihood of Model 7 and 8 is not directly comparable to that of other models. However, the marginal effects in the each pair of logit and probit models are very similar, suggesting that the robustness of the logit specification.

The above results are similar to those reported in the previous chapter for the static model. Models 1 to 8, while including the lagged dependent variable, do not consider the unobserved heterogeneity. Given the importance of accounting for unobserved effects in the dynamic model, both fixed effect models and random parameter models are also investigated.

The fixed effect version of Models 1 to 6 was estimated. With the presence of cohort dummy variables, the degree of freedom is reduced by 14, while the log likelihood is increased by 7 to 16 depending on the specific models. It should be noted that the fixed effects are treated as parameters in the model, and the finite T bias would be present for the model coefficients estimated using maximum likelihood method, although the bias might not be significant due to the relatively long sample period. Table 7-2 reported the results of two fixed effect model with log income and costs variables.

Comparing Model 13 and 14 with their pooled logit version of Model 5 and Model 6, the log likelihood increases by 10 and 15 respectively. The likelihood ratio test of fixed effects, with 14 degree of freedom, is not statistically significant for Model 13 but significant at 5% level for Model 14. This is not surprising as none of the coefficients for the cohort dummy in Model 13 is significant. In Model 14, the slope coefficient and marginal effect are larger for the younger cohorts, indicating a higher propensity of car ownership for the younger generations. Figure 7-1 illustrates the marginal effects of the cohort dummy variables, which has a clear linear trend except for the youngest cohorts. This result is similar to those reported for the linear models in Chapter 4 and 5 and those reported in Dargay and Vythoulkas (1999).

Table 7-2 Fixed Effect Models with Log Income and Cost Variables (t-stat in parenthesis)

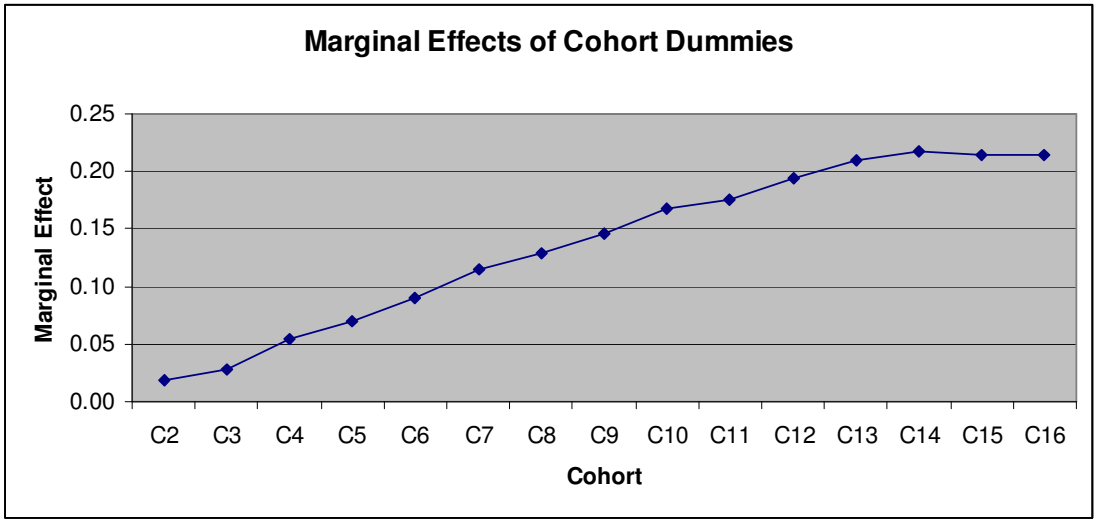
	Slope Coefficient				Marginal Effect			
	Model 13		Model 14		Model 13		Model 14	
LagY	1.6387	(8.93)	1.1311	(5.72)	0.3278	***	0.2263	***
LnInc	0.6329	(5.67)	0.2986	(2.61)	0.1266	***	0.0598	***
Child	-0.3762	(-2.71)			-0.0753	***		
Worker	-0.2835	(-4.02)			-0.0567	***		
HHSize	0.4648	(2.99)			0.0930	***		
HH2			-1.4670	(-2.52)			-0.2935	**
HH3			-1.9185	(-2.36)			-0.3839	**
HH4			1.2822	(2.76)			0.2565	***
HH5			0.7844	(1.36)			0.1569	'
HH6			1.3520	(2.79)			0.2705	***
HH7			0.8068	(1.37)			0.1614	'
HH8			0.9000	(1.57)			0.1801	'
Met	-0.8946	(-2.83)	-0.7237	(-2.27)	-0.1790	***	-0.1448	**
Rural	0.8815	(2.66)	0.5683	(1.70)	0.1763	***	0.1137	*
LnPrice	-0.5794	(-5.47)	-0.4621	(-3.81)	-0.1159	***	-0.0925	***
LnRunCst	-0.5412	(-4.92)	-0.5906	(-5.11)	-0.1083	***	-0.1182	***
Age	0.0467	(4.19)	0.0620	(4.76)	0.0094	***	0.0124	***
AgSq	-0.0554	(-5.62)	-0.0382	(-3.05)	-0.0111	***	-0.0077	***
C2	-0.0805	(-0.79)	0.0949	(0.90)	-0.0164	'	0.0186	'
C3	-0.0861	(-0.86)	0.1402	(1.31)	-0.0175	'	0.0272	'
C4	-0.0747	(-0.70)	0.2853	(2.33)	-0.0152	'	0.0537	**
C5	-0.0793	(-0.69)	0.3772	(2.76)	-0.0161	'	0.0699	***
C6	-0.0589	(-0.46)	0.5024	(3.23)	-0.0119	'	0.0905	***
C7	-0.0148	(-0.10)	0.6610	(3.76)	-0.0030	'	0.1148	***
C8	-0.0007	(-0.01)	0.7621	(3.94)	-0.0002	'	0.1294	***
C9	0.0182	(0.11)	0.8853	(4.10)	0.0036	'	0.1467	***
C10	0.0731	(0.38)	1.0314	(4.33)	0.0144	'	0.1670	***
C11	0.0416	(0.20)	1.1114	(4.24)	0.0083	'	0.1760	***
C12	0.0607	(0.26)	1.2655	(4.36)	0.0120	'	0.1936	***
C13	0.0807	(0.33)	1.4403	(4.50)	0.0159	'	0.2101	***
C14	0.0603	(0.23)	1.5981	(4.56)	0.0119	'	0.2168	***
C15	-0.1093	(-0.40)	1.6529	(4.27)	-0.0224	'	0.2136	***
C16	-0.2695	(-0.88)	1.7282	(3.98)	-0.0570	'	0.2139	***
Log Like'd	-65891		-65861					
Null LL	-81240		-81240					
Adj. LRI	0.1888		0.1891					

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Table 7-3 Short run elasticity derived from FE models with log income and price variables

	Model 13			Model 14		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.31	-0.28	-0.27	0.15	-0.23	-0.30
Middle	0.13	-0.12	-0.11	0.06	-0.10	-0.12
High	0.10	-0.10	-0.09	0.05	-0.08	-0.10

Figure 7-1 Marginal Effects of Cohort Dummies (Model 14)



One concern about the fixed effect model is the much smaller coefficient (and smaller t-stat) of the income variable, especially for Model 14, suggesting some explanatory power of income has been taken away by the cohort dummies. This would have some negative impacts on forecasts, as income is the exogenous variable that we can more easily control, while the fixed effects for the new cohorts are difficult to predict. Table 7-3 reports the short run elasticity of income, car purchase price and running costs for low, median and high income cohort, which have been calculated as $El = (\partial P/\partial x)/P$.

The short run income elasticity reported in Table 7-3 is lower than that derived from the pooled logit model (Model 5 and 6 in Table 7-1). While Model 14 seems to have better goodness of fit, the income elasticity is lower than those commonly found in the literature. Furthermore, the income elasticity becomes lower than the price elasticity in Model 14, inconsistent with earlier findings in Dargay and Vythoulkas (1999). On the other hand, the income and cost elasticity derived from Model 13 is more sensible, even though it has worse goodness of it.

One advantage of using dynamic model is the ability to capture long run relationship under equilibrium and the possibility of estimate long run elasticity. However, unlike the case of linear model, the long run elasticity can not be easily derived for dynamic discrete choice model. In the current study, we use Taylor expansion to derive approximate results. The formula used in the calculation of long run elasticity using

Taylor expansion is reported in Appendix 3 and the results for Model 13 and 14 are reported in Table 7-4.

Table 7-4 Long run elasticity derived from FE models with log income and price variables

	Model 13			Model 14		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.53	-0.48	-0.45	0.20	-0.32	-0.40
Middle	0.19	-0.17	-0.16	0.08	-0.12	-0.15
High	0.13	-0.12	-0.11	0.06	-0.09	-0.11

Table 7-5 Results of Random Parameter Model (t-stat in parenthesis)

	Mean of Param		Std Dv of Param		Marginal Effect	
ONE	-1.72791	(-0.88)	-0.00009	(-0.01)	-0.34581	'
LagY	1.49168	(8.10)	0.00001	(0.00)	0.29853	***
LnInc	0.79112	(6.55)	0.00000	(0.00)	0.15833	***
HH2	-0.97067	(-1.98)	0.00051	(0.02)	-0.19426	**
HH3	-1.11800	(-1.90)	0.00252	(0.03)	-0.22374	*
HH4	0.91559	(2.12)	0.00045	(0.02)	0.18324	**
HH5	0.22397	(0.44)	0.00100	(0.03)	0.04482	'
HH6	0.48834	(1.16)	0.00082	(0.04)	0.09773	'
HH7	0.05382	(0.10)	-0.00229	(-0.04)	0.01077	'
HH8	-0.03899	(-0.08)	-0.00304	(-0.03)	-0.00780	'
Met	-0.72991	(-2.43)	0.00104	(0.05)	-0.14608	**
Rural	0.47943	(1.46)	-0.00010	(-0.00)	0.09595	'
LnPrice	-0.60235	(-2.91)	-	-	-0.12055	***
LnRunCst	-0.13253	(-0.84)	-	-	-0.02652	'
Age	0.02671	(2.61)	0.00000	(-0.01)	0.00535	***
AgSq	-0.03006	(-2.84)	0.00000	(0.01)	-0.00602	***
Log Like'd	-65877					
Null LL	-81240					
Adj. LRI	0.1887					

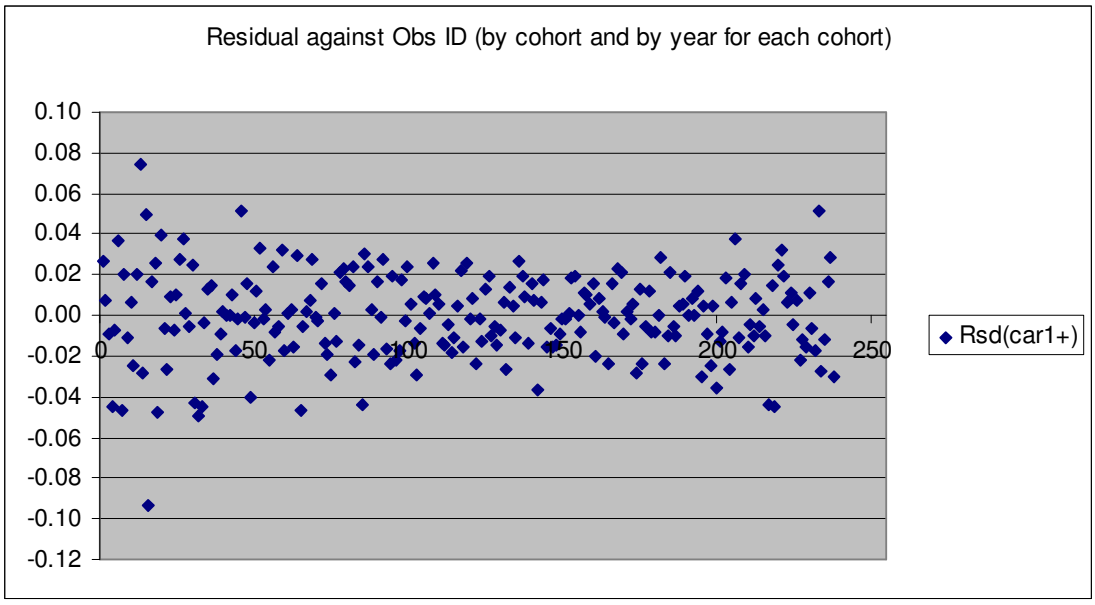
***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

The expansion point for the linear Taylor expansion corresponds to the proportion of households owning at least one car for the low, median and high income cohort. Comparing the results in Table 7-3 and 7-4, it shows that the long run elasticity is higher than short run elasticity by about 70% for Model 13 and by about 40% for Model 14 (low income household). Nevertheless, the long run income elasticity derived from Model 14 is still much lower than those reported in Dargay and Vythoulkas (1999), and is also lower than those derived from linear dynamic model in Chapter 5. These results are clearly unsatisfactory.

As discussed in Section 2, the fixed effect estimators are fixed *T* biased due to the incidental parameter problem. We then proposed to extend the random effect model to random parameter model, which relaxes the orthogonality assumption between the unobserved effects and the explanatory variables and accommodates lagged dependent variable. For the ownership model of one plus car, we estimate the random parameter version of Models 1 to 6 reported in Table 7-1. We report the results of the model with best fit in Table 7-5.

The model log likelihood of the random parameter model is -65877 and the adjusted likelihood ratio index is 0.1887, slightly lower than the fixed effect model. The residual plot (Figure 7-2) does not reveal any severe problems of auto-correlation and heteroskedasticity.

Figure 7-2 Residual plot of the Random Parameter model



The log variables of purchase price and running costs index are the same in any year for all cohorts, so their coefficients are treated as fixed parameters. The coefficients of other variables are assumed to be normally distributed. However, none of the estimated standard deviations of the random parameters (the square root of the diagonal elements of Γ in equation 16) are statistically different from zero. Regarding the mean of the random parameters, the constant term and log running cost parameter are not significant, while lagged dependent variable, log income, log purchase price, age parameters are all significant at 1% level. Only some of the household characteristic

variables and location variables have significant coefficients. Based on the mean of the random parameter, the short run and long run elasticity is reported in Table 7-6.

Table 7-6 Short Run and Long Run elasticity based on the mean of random parameters

	Short Run			Long Run		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.39	-0.29	*-0.06*	0.61	-0.46	*-0.10*
Middle	0.17	-0.13	*-0.03*	0.23	-0.18	*-0.04*
High	0.13	-0.10	*-0.02*	0.17	-0.13	*-0.03*

The running cost elasticity is not significant so is shown in italic. The short run income elasticity ranges from 0.13 for high income household to 0.39 to low income household, while the long run income elasticity ranges from 0.17 to 0.61. The purchase price elasticity varies between -0.10 and -0.29 in the short run while the range is between -0.13 and -0.46 in the long run. The income and price elasticity for high income households is about one third of that for low income households, and such difference is much bigger than that reported in Dargay and Vythoulkas (1999). This might suggest that the logit functional form adopted here better accounts for the impact of saturation.

Finally, it is worthy to provide some tentative explanations of why none of the random parameters has standard deviation significantly differently from zero. The first possibility is that the unobserved heterogeneity is no longer significant after the individual data are aggregated into cohorts. However, this argument is not supported by results of the fixed effect models, where the fixed cohort effects are significant for most of the cohorts under certain specification. Nevertheless, the improvement of the goodness of fit for the fixed effect model is not spectacular given the loss of degree of freedom (likelihood ratio test is significant at 5% but not at 1% level), which might suggest that the significance of heterogeneity is limited for cohort data. The second possibility is that the pseudo panel sample size might be too small to reliably estimate the distribution of parameters representing unobserved heterogeneity. Although the Family Expenditure Survey has thousands of observations each year, after they are aggregated into pseudo panel, we only have observations for 16 cohorts covering 19 years. This small sample size might not be enough to establish the distribution of the random parameters. The third possibility is that the assumption of normal distribution of parameters is inappropriate. However, alternative assumptions of uniform, triangular

and log-normal distribution all yield almost identical results so the assumption on parameter distribution is less likely to be the problem here.

7.3.2 Dynamic Model of Two plus Cars

The models of two plus cars are conditional on households owning the first car. As a result, it is based on the reduced pseudo panel constructed from survey observations of car owning households. As mentioned at the beginning of this section, the number of pseudo panel observations using in modelling is 206 after dropping the first observation for each cohort.

For the models of households with two or more cars conditional on owning the first car, specification search is similar to that described in the previous section and will not be reported here in details. The key results from the specification search are summarised as follows: firstly, the household characteristics variables that have statistically significant coefficients are the average number of children and people in work per household. Secondly, using compressed household location variables ("Met" and "Rural") leads to significant loss of model fit. Thirdly, the fixed effects are not significant and in some models their inclusion leads to loss of goodness of fit. Finally, there is no difference in log likelihood (model fit) whether the income and costs variables are in linear or log form.

Based on the above results, we identify the model with the best fit, which is reported in Table 7-7. To alleviate the problem of correlation between the explanatory variables and unobserved heterogeneity, random parameter (mixed logit) model is used. Table 7-7 reports the mean and standard deviation of the parameter of interest as well as the marginal effects evaluated at the weighted average of the explanatory variables.

The model log likelihood is -47152, and the Adjusted Likelihood Ratio Index is 0.1619. While no standard deviation of any random parameters is significantly different from zero, the mean of most parameters are significant and with expected sign. The means of two random parameters appear to have wrong sign: average number of children per household and log running costs. However, this appears to be the genuine results based on the data, as similar finding was obtained for static models in Chapter 6. Figure 7-4 is

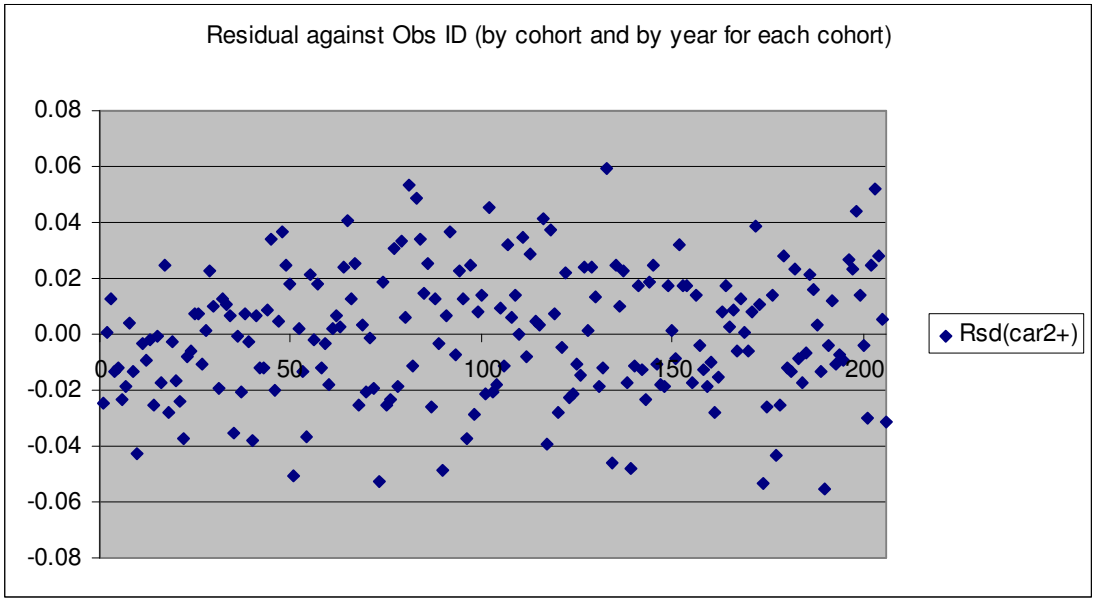
the residual plot of the Car 2+|1+ model, which does not indicate any problem of auto-correlation and heteroskedasticity.

Table 7-7 Random Parameter Model for Car 2+|1+ (t-Stat in parenthesis)

	Mean of Param		Std Dv of Param		Marginal Effect	
ONE	-7.52859	(-2.81)	-0.00006	(-0.01)	-1.53371	***
LagY	1.61840	(5.48)	0.00105	(0.04)	0.32970	***
LnInc	0.75088	(4.30)	-0.00008	(-0.06)	0.15297	***
Child	-0.12968	(-4.22)	-0.00027	(-0.04)	-0.02642	***
Worker	0.12459	(1.94)	-0.00004	(-0.01)	0.02538	*
AREA2	1.89829	(2.58)	-0.00096	(-0.02)	0.38672	***
AREA3	1.04462	(1.41)	-0.00030	(-0.01)	0.21281	'
AREA4	1.29271	(1.87)	-0.00133	(-0.04)	0.26335	*
AREA5	0.85025	(1.29)	0.00069	(0.02)	0.17321	'
LnPrice	-0.61564	(-2.39)	-	-	-0.12542	**
LnRunCst	0.33163	(1.78)	-	-	0.06756	*
Age	0.08985	(7.08)	0.00000	(0.02)	0.01830	***
AgSq	-0.09975	(-7.08)	-0.00001	(-0.05)	-0.02032	***
Log Like'd	-47152					
Null LL	-56288					
Adj. LRI	0.1619					

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Figure 7-3 Residual Plot of the Car 2+|1+ Model



The income and costs elasticity for low, median and high income cohorts are reported in Table 7-8. It should be noted that the overall income level for the car owning

households are higher, so these elasticity for Car 2+|1+ is evaluated at higher income than the model of one plus car. In general, the income and cost elasticity for the second (or more) car are higher than those for the first car. This result is as expected, since the ownership for second or third car in the household is less driven by necessity and is further away from saturation point. The elasticity differences between the low and high income cohorts are much smaller for 2+|1+ Car, mainly due to higher income and cost elasticity for high income cohort. In Table 7-8, both short run and long run elasticity have been reported, with the latter based on linear Taylor expansion described in Appendix 3. The long run income elasticity are higher than those in the short run by the range of 26% and 53%.

Table 7-8 Income and cost elasticity for model of Car 2+|1+

	Short Run			Long Run		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.64	-0.52	*0.28*	0.84	-0.69	*0.37*
Middle	0.55	-0.45	*0.24*	0.72	-0.59	*0.32*
High	0.47	-0.38	*0.21*	0.67	-0.55	*0.29*

7.4 Model with Saturation

As discussed in Chapter 6, there are two major advantages of using pseudo panel instead of cross sectional model in the empirical study of car ownership: the first is the consideration of dynamics, which has been investigated in the last three sections; the second is the modelling of saturation, which is the main focus here. In car ownership forecast model, saturation is an important concept. As pointed out by the Leitch Committee Report, "the accurate determination of the saturation level is of prime importance if the resulting forecasts are to command confidence. If the saturation level cannot be satisfactorily determined then the resulting forecasts are to that extent themselves unsatisfactory." (Department of Transport, 1978; cited in Whelan 2003, p79)

Saturation is a limit on the choices faced by decision maker, which may be reached but not exceeded. A model with saturation explicitly assumes that increasing income will bring car ownership levels closer to but never in excess of a saturation limit. Similar models that restrict range of possible choice fractions have been used under the name

of "Dogit" (Gaudry and Dagenais, 1979)[44]. While well established, there are outstanding issues with the interpretation and estimation of these saturation models, which will be addressed below.

7.4.1 Dogit Model

If the unconstrained discrete choice model is a binary logit model, the probability function of the corresponding model with saturation is commonly expressed as equation (22):

$$P = \frac{S \cdot \exp(x'\beta)}{1+\exp(x'\beta)} \qquad (22)$$

where S is the saturation level. However, it is not always possible to directly estimate (22), as such attempt failed miserably in the current study. Furthermore, S in (22) is simply a statistical parameter defining the upper asymptote, which is a parameter of no consequence in its own right (Button *et* al, 1982). We have yet to provide clear interpretation of S in the framework of Random Utility Model.

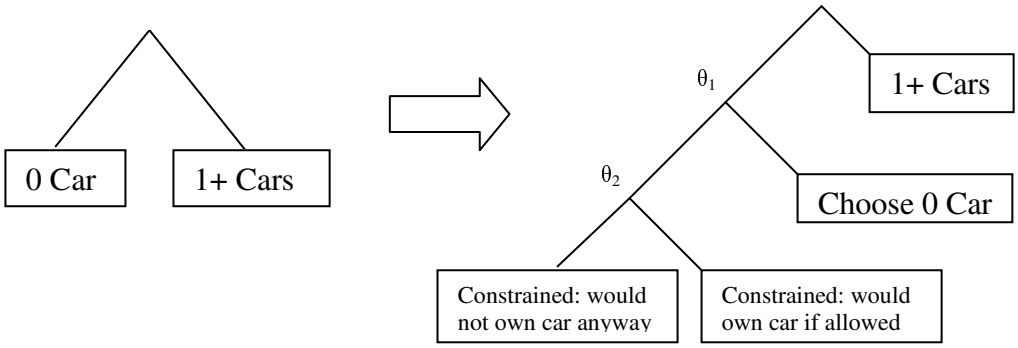
Saturation implies that some household are constrained not to own a car (captive to the alternative of zero cars). Reflected in the random utility model, the choice set faced by the decision maker[45] would have to be expanded to include new alternatives of "constrained choice". The difference in utility between the constrained and uncontained choice can then be used to infer level of saturation. However, it remains an issue how to estimate the random utility model with constrained choice. As shown in Daly (1999), by notionally separating the constrained choice set further into "voluntarily constrained" and "forcibly constrained", such model can be estimated using the conventional maximum likelihood method. The resulting model has a "tree logit" structure illustrated by Graph (b) in Figure 7-4.

When the choice set is formulated in the "tree-logit" form, Model (22) can be written in a way that is consistent with the random utility theory and can be estimated using conventional techniques. While the observable utilities of those choosing to own zero

[44] It is called Dogit Model because it dodges (avoids) the researcher's dilemma of choosing *a priori* between a format which commits to IIA restrictions and one which excludes them.

[45] The decision maker is an individual household in the micro survey. The probability model presented here refers to individual decision makers, and the corresponding pseudo panel model should be developed similarly as in Section 6.2.1 and 7.1.4. This is not shown in order to avoid duplication.

Figure 7-4 Choice Set when some decision makers are constrained not to own a car



or one plus car remain unchanged, the observable utilities of those constrained not to own a car would include an additional linear modifier: $S^*, (S^* \in R)$. As a result, the observable utilities for the four alternatives in Graph (b) are:

$V_{1+Car} = \beta' x$

$V_{Choose0} = 0$

$V_{forcibly\_constrained} = \beta' x + S^*$

$V_{voluntarily\_constrained} = S^*$

For identification, the scale parameters of the two lower nests (θ_1 and θ_2) have to be constrained to 1. In this case, the above nested logit model collapses into a multinomial logit model with four alternatives. For cross sectional model with discrete data, this model can be estimated using commercial software; for pseudo panel, the estimation routine is implemented in Gauss based on the mixed logit code for proportions data. More specifically, the probability of household choosing 1+ car can be expressed as:

$$P_{1+} = \frac{\exp(\beta' x)}{1 + \exp(\beta' x) + \exp(\beta' x + S^*) + \exp(S^*)} \qquad (23)$$

As researchers are not able to empirically distinguish household that choose not to own a car and those constrained not to own a car, the probability of household not owning a car has to be considered in aggregate:

$$P_0 = \frac{1 + \exp(\beta' x + S^*) + \exp(S^*)}{1 + \exp(\beta' x) + \exp(\beta' x + S^*) + \exp(S^*)} \qquad (24)$$

In (23) and (24), S^* reflects the impact of constraints on the probability of car ownership, whose sign, magnitude and statistical significance remained to be estimated. When the impact of such constraints on car ownership is negligible, then $S^* \to -\infty$, and the probability of household owning zero car is the same as when no constraints exists: $\lim_{S^* \to -\infty} P_0 = \frac{1}{1+\exp(\beta' x)}$. On the other hand, when the impact of such constraints on car ownership is extremely large, we have $S^* \to \infty$, and the probability of household owning zero car is close to one: $\lim_{S^* \to \infty} P_0 = 1$.

Model (24) is a special case of the Dogit Model proposed in Gaudry and Dagenais (1979)[46]. On the other hand, it is easy to show that the probability model of (23) is mathematically equivalent to equation (22), the empirical probability function commonly assumed for logit model with saturation. Rewriting equation (22) as:

$$P_{1+} = \frac{\exp(x'\beta)}{[1+\exp(x'\beta)]\cdot(1+\frac{1-S}{S})} = \frac{\exp(x'\beta)}{[1+\exp(x'\beta)]\cdot[1+\exp(\ln\frac{1-S}{S})]} \qquad (25)$$

and by noting $\ln\frac{1-S}{S} = S^*$ (26)

we would obtain the probability model of (23). Instead of directly estimating the non-linear term S in (22), we now estimate a linear term S^* in the exponential function in (23). As $S = \frac{1}{1+\exp(S^*)}$, it would satisfy $0 < S < 1$, representing the probability limit that can never be exceeded.

7.4.2 Empirical Results of Car Ownership Model with Saturation

The specification search is similar to that described in section 7.3.1 for models of one plus car and section 7.3.2 for models of two plus cars. While the consideration of saturation improves the goodness of fit, it does not change the comparative

---

[46] The Dogit Model has the probability distribution as: $p_i = \frac{\exp(V_i) + \theta_i \sum_j \exp(V_j)}{(1+\sum_j \theta_j)\sum_j \exp(V_j)}$. Model (24) is a binary Dogit with the following speficications: normalizing V_1 to zero, $\theta_1 = \exp(S^*)$ and $\theta_2 = 0$.

performance between models with different forms of explanatory variables and unobserved heterogeneity.

For model of **one plus car**, the model with the best fit is the fixed effect model, where the household characteristics are represented by the eight-way categorization of household types. While this model has the highest log likelihood after taking into account the degree of freedom, there are practical difficulties in using it for forecasts. One is the fixed effect for new cohorts in future years, which can not be estimated from existing data. The other difficulty is forecasting the split of households for the eight household types in each cohort, which, unlike the variables of household size and number of children/employed, can not be controlled using the forecasts published by the planning authority. Both issues would lead to additional uncertainty in the forecasts. Furthermore, the income elasticity implied by the fixed effect model is lower than the range identified in the literature, which is another cause of concern.

For these reasons, we also report the result of the "second best" model with alternative household characteristic variables. Initially, the model estimated was a random parameter (mixed logit) model; but as none of the random parameter has standard deviation significantly different from zero, the random parameter specification was abandoned and all parameters were treated as fixed. Table 7-9 reports the results of the fixed effect model and the "second best" pooled logit model for household owning at least one car, which will be used for forecasting in the next chapter.

The log likelihood of the fixed effect model with household type split is -65859, higher than the pooled logit model with average demographic statistics by about 40. The former has higher adjusted likelihood ratio index of 0.1889, which suggests better model fit after allowing for degree of freedom adjustment. As mentioned before, there are practical issues in applying the fixed effect model in forecasts, so it would be interesting to compare the forecasting results based on both models later on.

The coefficient for the linear modifier S^* is similar for both models. They translate to the saturation level of 0.9212 and 0.9437 respectively. However, the income elasticity differs significantly between these two models, with that for the pooled logit model being much higher. Table 7-10 compared the short run and long run income elasticity

Table 7-9 Forecasting model of one plus cars (t-statistic in parentheses)

	Slope Coefficient				Marginal Effect			
	Fixed Effect		Pooled Logit		Fixed Effect		Pooled Logit	
ONE			-4.8731	(-2.31)			-0.9628	**
LagY	1.0786	(4.45)	1.8888	(9.72)	0.2115	***	0.3732	***
LnInc	0.3945	(2.55)	1.1075	(6.52)	0.0774	**	0.2188	***
Child			-0.2550	(-1.74)			-0.0504	*
Worker			-0.2991	(-3.54)			-0.0591	***
HHSize			0.3950	(2.38)			0.0780	**
HH2	-1.3671	(-1.87)			-0.2681	*		
HH3	-2.0141	(-1.90)			-0.3950	*		
HH4	1.4763	(2.46)			0.2895	**		
HH5	1.0815	(1.44)			0.2121	'		
HH6	1.8374	(2.77)			0.3603	***		
HH7	1.2892	(1.58)			0.2528	'		
HH8	1.6144	(1.83)			0.3166	*		
Met	-0.9489	(-2.25)	-0.9420	(-2.60)	-0.1861	**	-0.1861	***
Rural	0.6772	(1.57)	0.9621	(2.51)	0.1328	'	0.1901	**
LnPrice	-0.5800	(-3.46)	-0.2662	(-1.17)	-0.1137	***	-0.0526	'
LnRunCst	-0.8184	(-4.30)	-0.2980	(-1.61)	-0.1605	***	-0.0589	'
Age	0.0919	(4.01)	0.0356	(2.99)	0.0180	***	0.0070	***
AgSq	-0.0583	(-3.16)	-0.0436	(-3.77)	-0.0114	***	-0.0086	***
S*	-2.4582	(-6.42)	-2.8195	(-6.90)	-	-	-	-
C2	0.1040	(0.92)			0.0204	'		
C3	0.1679	(1.45)			0.0329	'		
C4	0.3347	(2.44)			0.0656	**		
C5	0.4439	(2.82)			0.0870	***		
C6	0.5997	(3.24)			0.1176	***		
C7	0.8187	(3.72)			0.1605	***		
C8	0.9730	(3.85)			0.1908	***		
C9	1.1991	(3.87)			0.2351	***		
C10	1.4448	(3.90)			0.2833	***		
C11	1.5371	(3.86)			0.3014	***		
C12	1.7391	(3.94)			0.3410	***		
C13	1.9663	(4.04)			0.3856	***		
C14	2.1536	(4.10)			0.4223	***		
C15	2.2183	(3.91)			0.4350	***		
C16	2.3610	(3.73)			0.4630	***		
Log Like'd	-65859		-65900					
Null LL	-81240		-81240					
Adj. LRI	0.1889		0.1887					

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Table 7-10 Short run and long run income elasticity of one plus car model

	Short Run Income Elasticity		Long Run Income Elasticity	
Income	Fixed Effect	Pooled Logit	Fixed Effect	Pooled Logit
Low	0.198	0.550	0.238	0.925
Middle	0.082	0.240	0.111	0.413
High	0.065	0.191	0.069	0.252

for these two models. The impacts of different income elasticity on car ownership forecast will be examined in the next chapter.

Regarding the model of household owning **two or more cars** conditional on owning the first car, specification search shows that fixed effect models do not have better goodness of fit. While using the five-area household location split improves model fit, there is no significant loss of fit when using average demographic statistics rather than eight-way household type split if degree of freedom is taken into account. The average household size variable is not significant and subsequently dropped, so the household characteristics are described by the average number of children and people in work per household. This leads to the model of best fit reported in Table 7-11. It should be noted that the model initially estimated was a random parameter model; as the standard deviations of the random parameters were not significantly different from zero, all parameters are treated as fixed in the final model.

Table 7-11 Model of Car 2+|1+ (t-stat in parenthesis)

	Slope Coeff		Marginal Effect	
ONE	-10.4365	(-3.08)	-2.1587	***
LagY	2.3361	(5.23)	0.4832	***
LnInc	1.0649	(4.63)	0.2203	***
Child	-0.1362	(-3.67)	-0.0282	***
Worker	0.1844	(2.26)	0.0381	**
AREA2	2.2701	(2.77)	0.4695	***
AREA3	1.1823	(1.45)	0.2446	'
AREA4	1.6324	(2.11)	0.3376	**
AREA5	1.0771	(1.44)	0.2228	'
LnPrice	-0.6078	(-1.88)	-0.1257	*
LnRunCst	0.6191	(2.42)	0.1280	**
Age	0.0769	(5.15)	0.0159	***
AgSq	-0.0840	(-5.08)	-0.0174	***
S*	-0.7891	(-3.07)	-	
Log Like'd	-47147			
Null LL	-56288			
Adj. LRI	0.1621			

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Similar to the unconstrained Car 2+|1+ model reported in Table 7-7, there are two parameters with unexpected sign. The coefficient of average number of children in the household is negative and significant, but it might be due to the correlation between that variable and the average number of people in work. The latter is significant at 5%

level, with marginal effects on the conditional choice probability of 0.038. While the coefficient for log of real purchase price is negative and significant, that for the log of real running costs is significant but of wrong sign. This result was previously identified for the static and unconstrained dynamic model of Car 2+|1+ and might be caused by the concurrent substantial rise of car running costs and ownership of two plus cars in the second half of 1990s. In terms of household location, if the proportions of households living in metropolitan and rural areas (Area type 2 to Area type 4) increase at the expense of that in Greater London (the base case of Area type 1), the conditional probability of household owning two or more cars would also increase. Finally, the coefficients for the average age of household head and age square (dividing by 100) are positive and negative respectively, indicating a peak of car ownership during the household life cycle.

The estimated linear utility modifier S^* is -0.7891, implying a saturation level of 0.6876. We have also calculated the short run and long run income and costs elasticity for cohorts with low, median and high income level, which is reported in Table 7-12, The income and purchase price elasticity are higher than those for models of one plus car, which is as expected. On the other hand, the running cost elasticity is shown in italic due to its unexpected sign.

Table 7-12 Income and cost elasticity of Car 2+|1+

	Short Run			**Long Run**		
Income	Income Elasticity	Price Elasticity	Running Cost Elasticity	Income Elasticity	Price Elasticity	Running Cost Elasticity
Low	0.95	-0.54	*0.55*	1.23	-0.70	*0.72*
Middle	0.78	-0.45	*0.45*	0.94	-0.53	*0.54*
High	0.62	-0.35	*0.36*	0.68	-0.39	*0.39*

7.5 Conclusion

In this rather long chapter, we tackle two important issues that motivate the estimation of non-linear pseudo panel model in the first place: the consideration of dynamics and saturation in the car ownership choice. Since the 1980s, there have been a growing number of researchers that recognised the importance of dynamics and applied it in transport studies (e.g. Hensher and Wrigley, 1986; Kitamura, 1990; Mears et al. 1990; Goodwin et al. 1990; Goodwin, 1997; Long, 1997). In the current study, we focus on the methodological aspects of the nonlinear dynamic models. In particular, the first two

sections deal with the development and consistent estimation of the dynamic discrete choice pseudo panel model. We have shown in the previous chapter that the utility function of the pseudo panel model is a direct transformation from its cross-sectional counterpart, and if the measurement error can be ignored, these two types of models have the similar probability function, albeit with different scale (random term). This result facilitates the discussion in this chapter, where the behaviour models are more conveniently developed for individual decision maker before transforming into pseudo panel model.

In developing a dynamic car ownership model, we follow a general to specific approach, starting from a structural model with three forms of true state dependence. The general model encompasses three specific models: standard state dependence model, propensity dependence model and dynamic optimisation model. In propensity dependence model, the lagged effect is captured by the previous tendency to select a state (choice), which is unobservable and will lead to additional uncertainty in forecasts. Model of dynamic optimisation, despite its promise of enriching choice dynamics, requires a fundamental shift of car ownership model from holding model to transactions model. As a result, the standard state dependence model is chosen as the preferred model of dynamic car ownership choice.

The estimation of the nonlinear dynamic pseudo panel model is examined theoretically and empirically. We first review the various fixed effect, random effect and semi-parametric estimators proposed for the genuine panel data in the literature; subsequently, the fixed effect model and random parameter model of pseudo panel are proposed for the current study. In the empirical section, separate models have been estimated for households owning at least one car and those owning two or more cars conditional on the ownership of the first car. Both fixed effect models and random parameter models are tested, although the standard deviations of the random parameters are not significantly different from zero in all models.

The other important issue investigated in this chapter is the specification and estimation of car ownership model with saturation, which is a key consideration for the use of discrete choice method. In the framework of random utility model, saturation implies that some households are constrained not to own a car. Accordingly, the choice set

faced by the decision makers has been expanded to include the new alternatives of "constrained zero cars". To facilitate estimation, the model is specified with a "tree logit" structure and instead of directly estimating the saturation level S (a nonlinear term in the probability function), we estimate a linear modifier S^* in the utility function. The constrained dynamic model is subsequently implemented in the empirical study of car ownership. The estimated models of Car 1+ and Car 2+|1+ will then be used to forecast car ownership in Britain to year 2021, which will be the subject of next chapter.

Chapter 8 Car Ownership Forecasts

In the last four chapters, a range of car ownership models have been estimated using the pseudo panel dataset constructed from the Family Expenditure Survey. The systematic specification search determines the models with best fit, which will be used in car ownership forecasting here. The forecasting models include both linear and nonlinear models so their performance can be compared. In the current study, the geographic area covered is limited to Great Britain (as opposed to the United Kingdom) to be consistent with the National Road Traffic Forecasts (NRTF) and the National Transport Model maintained by the Department for Transport. The forecasting period is between 2001 and 2021, since more detailed household projection data are only available up to year 2021.

As all the empirical models are estimated using pseudo panel data, which are average statistics of cohort sample, it is easier to obtain aggregate measures such as total car stock compared to cross sectional models. Unlike the latter, it is not necessary to use the more complicated techniques such as prototypical sample enumeration (Daly and Gunn, 1985; Whelan, 2003; Whelan, 2007). However, it is still a challenging task to derive the cohort based household characteristics in future years using the available planning data. More specifically, it is important to separate the age effects and time trend effects (similar to the 'life cycle effects' and 'generation effects' in Dargay and Vythoulkas, 1999) on income and other characteristics over the life cycle. This issue will be discussed in Section One, which deals with projection of explanatory variables and other relevant variables over the forecasting period. Section Two uses the projected input data and the model parameters estimated in the previous chapters to generate forecasts. The forecasts results have been validated to the observed data and compared to those obtained from other studies. A number of scenario tests have also been carried out to ensure the models have the appropriate sensitivities. Section Three is a brief conclusion.

8.1 Projection of Explanatory Variables

It is necessary to first establish the age profile of the existing and new cohorts over the forecasting period. We decide to drop data points when household head is aged over 100, which also leads to the exclusion of the oldest cohort in the dataset (born between 1901 and 1906) from the forecasting model. On the other hand, five new cohorts have been introduced over the period, with the youngest born between 2001 and 2006. As a result, the model involves 20 cohorts in total over the forecasting period.

8.1.1 Forecast Assumptions

For each of the twenty cohorts in the period between 2001 and 2021, one has to make projections of explanatory variables in the econometric models and two other variables: number of households and 'multiple-car factor' for households with two or more cars. The explanatory variables include household real disposable income, average household demographic statistics, split of households between the eight household types, split of households between location types and aggregate real purchase price and car running costs index. The input data in 2000 (Year 0) are estimated based on Family Expenditure Survey data or backcast of 2001 census data. The future year growth assumptions are derived from social economic forecasts published by various sources.

Regarding the number of households in the 20 cohorts, we make use of the census product from Office of National Statistics, "Focus on Family" (ONS, 2005), which contains data on the number of families based on the 14 age bands of family reference person in 2001. By further taking into account the number of one person household in different age groups, it is possible to derive the number of household for all cohorts in 2001. ODPM (1999)[47] and Scottish Executive (2002) provide projections of household in England and Scotland and are used to derive the growth rates of household number by different age bands of household representatives.

Regarding household real disposable income, the base year (Year 2000) data is obtained from Family Expenditure Survey. The income growth is assumed to be in line with the growth of Gross Domestic Product (GDP), which however has to be adjusted

[47] While more recent household projection data have been published, they do not provide the growth rate by age of household representative so can not be used.

downward to account for the increasing number of household in each cohort. We use the observed real GDP growth between 2001 and 2006, which is obtained from Treasury Weekly Economic Indicators Databank (Treasury, 2007). From 2007 onwards, the GDP is assumed to grow at 2.25% per annum, the same rate used in Department for Transport's National Road Traffic Forecasts (1997), National Transport Model (Whelan, 2003; 2007) and 10 year plan (DETR, 2000).

The base year estimates of household size and average number of children and person in work per household by cohorts are obtained from Family Expenditure Survey. The publication by Government Actuary's Department on projected populations by age (GAD, 2003) is used to calculate the growth rate of household size and average number of children per household (taken into account the change of household numbers). Regarding the average number of working persons per household, we assume a constant labour market participation rate of 74.6% of the adult population (Treasury, 2007), so the workforce growth is entirely driven by population change.

It is not possible to project the change of location split by cohorts. As a result, the base year location split estimated from Family Expenditure Survey is assumed to be unchanged over the forecasting period. Regarding the real car ownership costs index, we assume the car purchase price falls by 0.37% per annum and the car running costs remain constant. These assumptions are also consistent with those in the National Transport Model and the 10 year plan.

8.1.2 Generating Projections of Input Variables

The analysis in Chapter 3 reveals that for pseudo panel data, household characteristics such as income go through a "hump" shape life cycle peaking at the age of late 40s; furthermore, at a given age, households in younger cohorts tend to have higher income than those in older cohorts. To derive sensible projection of the input variables, one should separate the age effect and time trend effect. The current study develops a sub-model of input projection, which includes 81 overlapping age bands and explicitly separates these two effects. This sub-model is implemented in three steps:

1. Estimating the base year figures for the relevant variables for 81 overlapping age bands of household head, e.g. those aged 15-19, 16-20, 17-21…94-98, 95-

99. The data sources include census and Family Expenditure Survey, and because the original data are for non-overlapping age groups (15-19, 20-24, 25-29…), method of interpolation is used to obtain estimates for all 81 age bands. This stage isolates the age effect cross cohorts.

2. For each of the 81 age band, forecast the future year figures based on standard growth assumption described in the previous sub-section. Different growth rates are applied to different cohorts whenever it is possible. This stage introduces the time trend effect.
3. The first two steps have produced a matrix of 21 rows by 81 columns (21 years for 81 age bands) for each input variable. Within each matrix, identify the twenty cohorts by the age of the household head. For example, in 2001, age band 16-20 corresponds to cohort whose head is born between 1981 and 1985 (Cohort ID F5); age band 21-25 is cohort born between 1976 and 1980 (ID F6). In 2002, it is age band 17-21 that refers to cohort F5 and age band 22-26 refers to cohort F6. Similarly, age band 36-40 refers to cohort F5 and age band 41-45 refers to cohort F6 in 2021. Extract the appropriate cells for each of the 20 cohorts from the 21x81 matrix and arrange them by cohort and year, we obtain the projection of input variables that can be used in the car ownership forecasting models.

The above method is used to generate projection for most of the input variables. However, an alternative approach has to be adopted regarding the split of eight household types, because it is not possible to obtain the appropriate growth rate required in the second step of the projection model and it is not satisfactory to assume that there is no change of household type split within cohort over time. The alternative approach involves assigning the observations in the original pseudo panel dataset into a 69 by 20 matrix. The 69 rows cover cohorts aged between 19 and 87, and the 20 columns correspond to the 20 cohorts. At a certain age (in one particular row), there are a number of pseudo panel observations belonging to different cohorts, which gives the growth rate between generations (younger and older cohorts). To dampen down noise, we actually use the average growth rates of cohorts with similar age (within 5 years difference) to project the future year values. These future year figures are contained in different cells of the 69x20 matrix and have to be extracted and re-arranged by cohorts

and years. A final adjustment is made to ensure that the proportions of the eight household types sum to 100%.

8.1.3 Checking of Projection Results

Before using the projected values of the explanatory variables in forecasts, it is necessary to check whether the projections are sensible. One way of checking is to examine the life cycle profile of the explanatory variables in the forecasting period. The variables of household real disposable income and average household size are used as example here.

Figure 8-1 shows the profile of the income variable for nine selected cohorts, covering age between 19 and 97. The income profile has a "hump shape" life cycle with the peak occurs when the head of household is in his early 50s. Furthermore, the younger cohorts always have higher income compared to the older ones at the same age, reflecting the so-called "generation effects". At the same age, the typical income difference between the two adjacent cohorts in Figure 8-1 is about 17%. Such difference is consistent with the typical income growth assumption of around 1.6% per annum (taking into account smaller household size and larger household number) and the 10 years gap between those cohorts. Comparing Figure 8-1 with Figure 3-5 in Chapter 3, one can see that the observed data and the projected data have broadly similar profile, although the former obviously has much more noise.

Figure 8-2 shows the life cycle of average household size for nine cohorts at different ages. The life cycle pattern is similar to that of household income, although for household size it peaks at the much earlier age of late 30s. This result is consistent with that obtained from the Family Expenditure Survey data. The generation effect is also present, and the younger cohorts always have smaller household size at the same age. It is consistent with the observed demographic change and reflects the household size growth assumptions of between -0.04% and -0.37% per annum. Besides income and household size, similar checks have been carried out for other input variables and the results are found to be satisfactory.

Figure 8-1 Projected weekly disposable income: Profile by age of household head

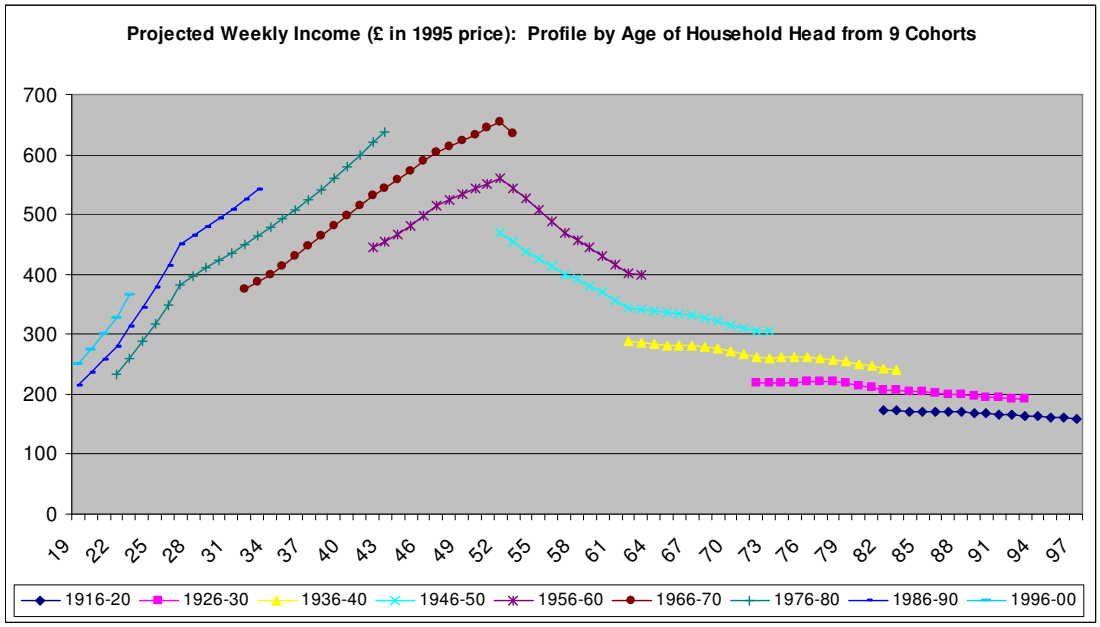
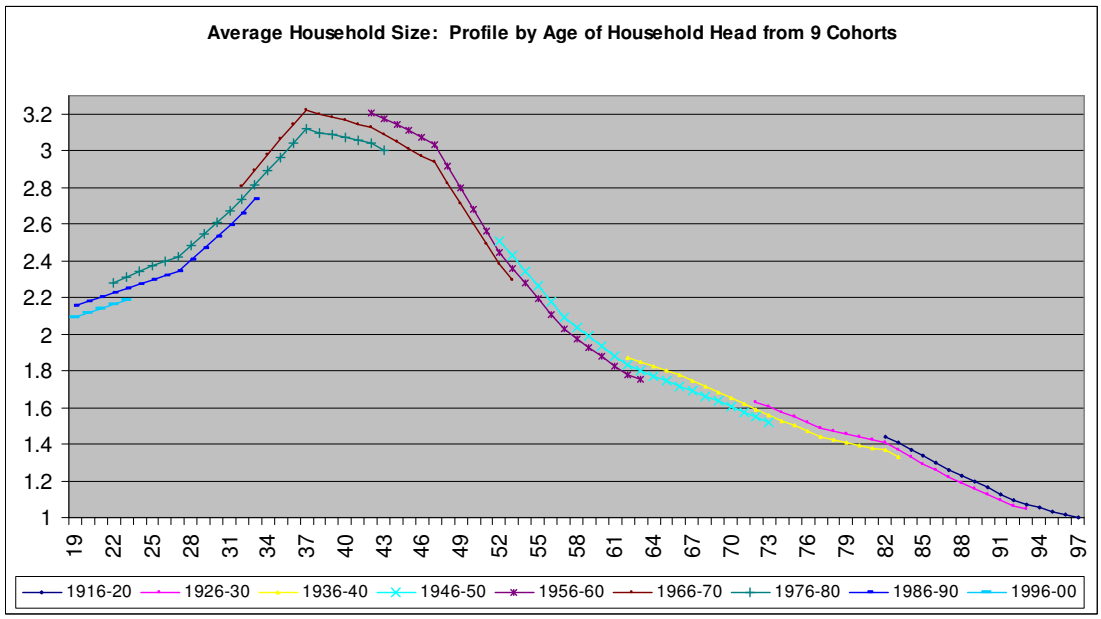


Figure 8-2 Projected average household size: profile by age of household head



8.2 Car Ownership Forecasts and Model Performance Evaluation

After projecting the future values of the explanatory variables in a sub-model, we are ready to apply the econometric models to generate car ownership forecasts. A number of econometric models have been used and the results are compared to the observed data between 2001 and 2006 as well as forecasts from other published studies. Several

scenario tests have been carried out to ensure that the forecasting models have the right sensitivity.

8.2.1 Selection of Econometric Models

A range of econometric models have been reported in Chapter 4 to 7, and it is important that the most appropriate ones are selected for forecasting. Some models are not suitable for forecasting purpose. For example, the linear models based on Weighted Least Square Estimator assume that economic relationship between the dependent variable and explanatory variables is linear for individuals in the micro survey. For such models, the linear transformation is done on individual data and the pseudo panel observations are the average of the transformed data (e.g. average of log income rather than log of average income). Besides the difficulty in supporting the assumption of linear relationship between car ownership and the explanatory variables for individual household, the fundamental problem is the impossibility to derive future values of the average transformed variables. As a result, the linear models used in forecasts are those assuming linear economic relationship at cohort level.

Furthermore, some models are more suitable for analytical purpose rather than forecasting purpose. The most notable examples are semi-parametric models and conditional logit models discussed in Chapter 6 and 7. After eliminating the nuisance parameters, the marginal effects of these models are not defined, which renders them useless in forecasting. For this reason, no such empirical models have been estimated. Another class of models that is less suitable for forecasting purpose is heterogeneous models. As shown in Baltagi and Griffin (1997) and Baltagi et al. (2003), "simplicity and parsimony in model estimation offer more plausible estimates and better forecasts"; homogenous models offer better out-of-sample forecasting performance, even though the hypothesis of homogeneity is formally rejected in the statistical tests. Furthermore, when using the random parameter model in forecasts, the random parameters themselves have to be simulated, so the likely outcome is more noise in the forecasts. For this reason, the heterogeneous models of individual cohorts and random parameter models are not taken forward in the forecasts[48].

[48] Actually, the estimated standard deviation of the random parameters in the mixed logit models are not statistically different from zero, so there are obviously no benefits in using these models.

After some careful selection, five dynamic models[49] have been selected to generate four sets of forecasts in order to compare the impacts of functional forms and choices of explanatory variables on the forecasts results. They include a linear fixed effect model (noting as Model L1 for convenience), a linear restrictive fixed effect model (Model L2), a fixed effect Dogit model for Car 1+ (Model D1), a pooled Dogit model for Car 1+ (Model D2) and a pooled Dogit model for Car 2+|1+ (Model D3). The parameters of these five models are reproduced in Table 8-1.

Model L1, D1 and D3 are the ones that were found to have the best goodness of fit in the specification search, and were reported in Table 5-4, Table 7-9 and Table 7-11 respectively. However, there are two issues when using Model L1 and D1 in forecasts. Firstly, both models are fixed effect models, and there are uncertainty regarding the assumptions of fixed effects for new cohorts. Secondly, both models use the split of eight household types as explanatory variables; however, there are no aggregate planning data on which the projection of these variables can be based. Regarding the first issue, it seems most appropriate to assume that the fixed effects for the new cohorts are the same as the youngest one in the sample, as the linear trend stops and the fixed effects become levelled for the younger cohorts (See Figure 5-6 and 7-1). Regarding the second issue, we have to rely on the alternative methods described in the previous section to identify growth from past trends.

As the solutions to these two issues are not entirely satisfactory, it seems worthy to generate forecasts using other models. More specifically, the alternative linear model restricts the cohort fixed effects to be linear (Model L2), and the alternative nonlinear model of Car 1+ ignores cohort fixed effect (Model D2). In both models, the household characteristics are represented by the average number of children, worker and household size rather than split of household types. By comparing the four sets of forecasts, one can establish whether the results are robust and how sensitive they are to model specification.

[49] As the initial condition for new cohorts has to be estimated using the parameters of static models, there are actually 10 econometric models in total that have been used in forecasting.

Table 8-1 Parameters of Econometric Models Used in Forecasts

	L1: FE	L2: Rst FE	D1: Car1+ FE	D2: Car1+ Pooled	D3: Car2+|1+
Constant		-3.0926		-4.8731	-10.4365
LagY	0.1889	0.3447	1.0786	1.8888	2.3361
LnInc	0.1894	0.2901	0.3945	1.1075	1.0649
Child		-0.1740		-0.255	-0.1362
Worker		0.0137		-0.2991	0.1844
HHSize		0.1637		0.395	
HH2	-0.1090		-1.3671		
HH3	-0.3393		-2.0141		
HH4	0.4702		1.4763		
HH5	0.4802		1.0815		
HH6	0.5201		1.8374		
HH7	0.9149		1.2892		
HH8	0.7286		1.6144		
Met	-0.1612	-0.3235	-0.9489	-0.942	
Area2					2.2701
Area3					1.1823
Area4					1.6324
Area5 (Rural)	0.1942	0.3640	0.6772	0.9621	1.0771
LnPrice		0.1855	-0.58	-0.2662	-0.6078
LnRunCst	-0.0958	-0.0260	-0.8184	-0.298	0.6191
Age	0.0306	0.0197	0.0919	0.0356	0.0769
AgSq	-0.0002	-0.0001			
AgSq/100			-0.0583	-0.0436	-0.084
Cohort		0.0451			
C1	-1.7403				
C2	-1.6816		0.104		
C3	-1.6537		0.1679		
C4	-1.5985		0.3347		
C5	-1.5547		0.4439		
C6	-1.4920		0.5997		
C7	-1.4019		0.8187		
C8	-1.3214		0.973		
C9	-1.2043		1.1991		
C10	-1.1352		1.4448		
C11	-1.0751		1.5371		
C12	-1.0067		1.7391		
C13	-0.9314		1.9663		
C14	-0.8541		2.1536		
C15	-0.8214		2.2183		
C16	-0.8104		2.361		
OUT	-0.1554	-0.1720			
Saturation			0.921	0.944	0.688
Adj R^2	0.994	0.992			
Adj LRI ($\overline{\rho}^2$)			0.1889	0.1887	0.1621

8.2.2 Forecasting and Validation

For linear model, the dependent variable is the average number of cars per household, so the total car stock can be easily obtained by multiplying the fitted dependent

variable by the household numbers in each cohort and summing over all cohorts. For nonlinear model, we estimate the probability of household owning at least one car (P_{1+}) and owning two or more cars conditional on owning the first one ($P_{2+|1+}$). The unconditional probability of household owning two or more cars are the product of $P_{2+|1+}$* P_{1+}. When the discrete choice model is a multinomial logit model with a constant term, first order condition ensures that these probabilities are the unbiased estimates of proportions of households owning certain number of cars. It thus follows that the proportion of household owning *exactly* one car is (P_{1+} - $P_{2+|1+}$* P_{1+}) and the proportion of household owning two plus cars is $P_{2+|1+}$* P_{1+}.

For those with two or more cars, one has to estimate the average number of cars in the household, or the so-called 'multiple-car factor' (noted as F, $F \geq 2$). The base year values of multiple-car factors (F_{c0}) are derived using the Family Expenditure Survey data. The long term growth rate of F is assumed to be 0.10% per annum, calculated using FES data over a 10 year period. Table 8-2 shows the assumed average number of cars in multiple-car household for six age bands in five years[50].

Table 8-2 Multiple-car factor used in forecasting

	16-19	20-24	25-44	45-64	65-74	75+
2001	2.002	2.022	2.156	2.298	2.065	2.002
2006	2.013	2.033	2.167	2.310	2.076	2.013
2011	2.023	2.043	2.178	2.322	2.086	2.023
2016	2.034	2.054	2.190	2.335	2.097	2.034
2021	2.044	2.065	2.201	2.347	2.108	2.044

For every year, the total number of cars is then calculated by multiplying the total number of households by the proportions of car owning households for each cohort, and summing over all cohorts:

$$TA_t = \sum_c HH_{ct}[(P_{1+}^{ct} - P_{2+|1+}^{ct} \cdot P_{1+}^{ct}) + (P_{2+|1+}^{ct} \cdot P_{1+}^{ct}) \cdot F_{ct}]$$

$$= \sum_c [HH_{ct} \cdot P_{1+}^{ct} + HH_{ct} \cdot P_{2+|1+}^{ct} \cdot P_{1+}^{ct} \cdot (F_{ct} - 1)]$$

where, TA_t = Total number of cars in year t;

HH_{ct} = Total number of household for cohort c in year t;

[50] To obtain multiple-car factor for all cohorts, we follow a process similar to the sub-model of input projection, which involves calculating the future year factors in a 21 by 81 matrix.

F_{ct} = Average number of cars in multiple car household for cohort c in year t.

Initial forecasts are produced using the model parameters reported in Table 8-1. These results are then compared to the observed car stock in Great Britain between 2001 and 2006. The total car stock are calculated from Transport Statistics Bulletin 'Vehicle Licensing Statistics' (DfT, 2005; 2006a) and 'Vehicle Exercise Duty Evasion' (DfT, 2006b), with the latter providing estimates of unlicensed car stock. The total number of cars includes private cars (whether owned by individuals or companies) but excludes "non-cars private light goods vehicle". Table 8.3 compared the four sets of forecasts against the observed car stock.

Table 8-3 Observed Total car stock *vs.* forecasting results (000s)

	Observed	L1	L2	D1+D3	D2+D3
2001	24,951	24,688	24,008	24,800	24,645
2002	25,623	25,235	24,369	25,300	25,085
2003	25,897	25,736	24,957	25,822	25,570
2004	26,502	26,047	25,672	26,181	26,133
2005	27,020	27,052	26,439	27,058	26,725
2006	27,648	27,845	27,106	27,546	27,167

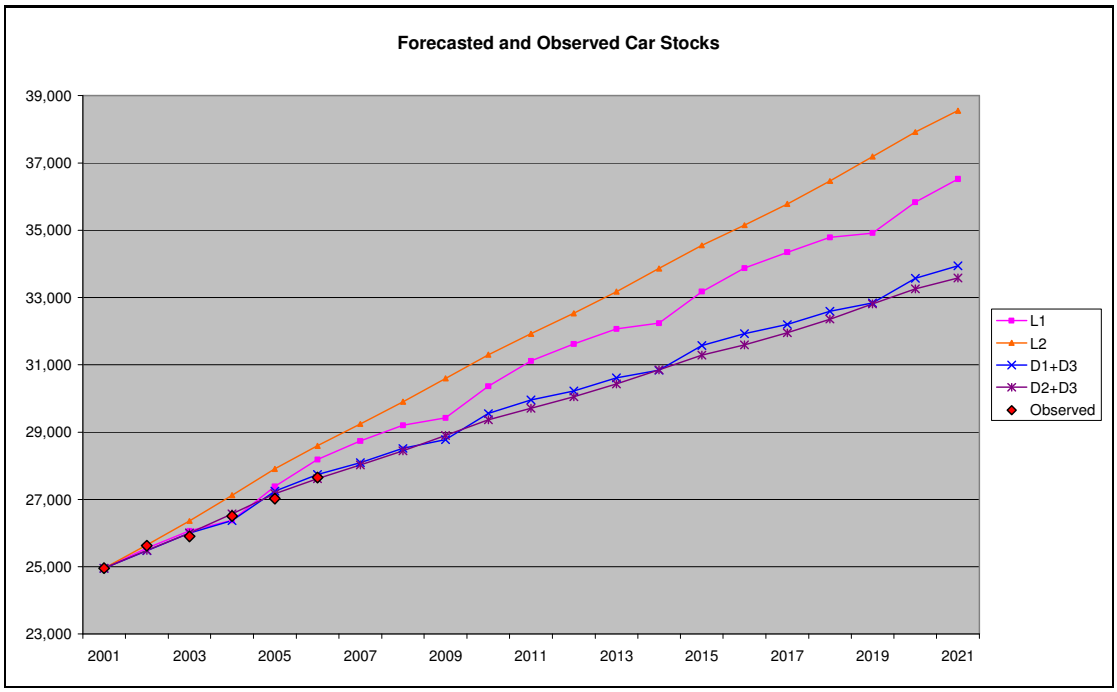
Overall, the results from four models are all close to the observed figures, which should give confidence to our forecasting model. However, it should be beneficial to apply a small adjustment to the model parameters so that all 2001 forecasts are validated against the observed figure. Such adjustment also makes it more transparent in the comparison of the four models. The following changes are subsequently made:

- All cohort fixed effects are increased by 0.011 in Model L1;
- The constant term is increased by 0.0396 in Model L2;
- The saturation level is increased from 0.921 to 0.927 in Model D1;
- The saturation level is increased from 0.944 to 0.955 in Model D2.

The above adjustments ensure the 2001 forecasts match the observed number of cars of 24,951 in all four sets of forecasts. The post validation results are illustrated in Figure 8-3, where the observed car stock is represented by 'red diamond' points. The forecasts based on nonlinear model D1 and D2 (both with Car 2+|1+ model of D3) closely match the observed figures between 2001 and 2006. On the other hand, the linear fixed effect

model L1 over-predicts the car stocks by 1.9% in 2006, and the other linear model L2 over-estimates by a higher 3.4%.

Figure 8-3 Observed Total Car Stocks and Forecasts from four Models



The similar results from model D1 and D2 suggest that the impacts of using different household characteristics variables on forecasts are small; also, assuming the new cohorts have the same fixed effects as the youngest one in the sample leads to similar outcome as the pooled logit model. On the other hand, both linear models generate higher forecasts, and the difference becomes larger in future years. Because linear models do not explicitly account for saturation, such results are in line with expectation. Finally, restricting the fixed effect to be linear leads to high car ownership in the new cohorts and hence the much higher aggregate forecasts.

For nonlinear models, it is also possible to compare the forecasted proportions of households owning cars to the observed values. The observed data are available up to year 2004, obtained from the latest Transport Statistics Great Britain (DfT, 2006c). It should be noted that the reported data are the proportions of households with regular use of *cars and vans*, while our forecasts refer to cars only. The comparisons are presented in Table 8-4.

Table 8-4 Proportion of Households with zero, one and two plus cars

	Zero Car			One Car			Two or more Car		
	Obs	D1+D3	D2+D3	Obs	D1+D3	D2+D3	Obs	D1+D3	D2+D3
2001	26%	28%	27%	45%	46%	46%	29%	27%	27%
2002	26%	27%	27%	44%	45%	45%	30%	28%	28%
2003	26%	27%	27%	44%	45%	45%	30%	28%	28%
2004	25%	27%	26%	44%	44%	45%	31%	29%	29%

The first impression from Table 8-4 is that our forecasts have higher proportions of household with no cars and lower proportions with two plus cars. The actual result is that the forecasted proportions of Car 1+ and Car 2+|1+ are both lower than the reported figures (note the proportion of household with exactly one car is calculated as P_{1+} - $P_{2+|1+}$* P_{1+}). This is what we should expect, as our models do not include private vans.

8.2.3 Forecasts Evaluation and Sensitivity Test

Beyond year 2006, when no observed car ownership data are available, the forecasts are evaluated using alternative methods. The analyses include comparison to other published studies, examination of car ownership profiles by age of household head and various sensitivity tests.

The other studies used for comparison include National Road Traffic Forecasts (NRTF, 1997), car ownership model supporting the influential RAC report "Motoring Towards 2050" (RAC 2002a; 2002b), and car ownership sub-model in the UK Department for Transport's National Transport Model (Whelan, 2003 and Whelan, 2007). Table 8-5 compares the four sets of forecasts in the current studies with the above sources.

Table 8-5 Forecasts Comparison: current studies *vs.* published studies (millions)

Year	L1	L2	D1+D3	D2+D3	NRTF (1997)	RAC (2002b)	Whelan (2003)*	Whelan (2007)*
2001	24.95	24.95	24.95	24.95	25.18	25.18	28.12	25.63
2006	28.18	28.59	27.74	27.62	n.a.	n.a.	30.28	28.59
2011	31.12	31.92	29.95	29.71	28.88	28.88	32.66	30.84
2016	33.87	35.15	31.92	31.59	n.a.	n.a.	34.48	32.71
2021	36.52	38.56	33.94	33.58	31.77	32.26	36.08	34.26

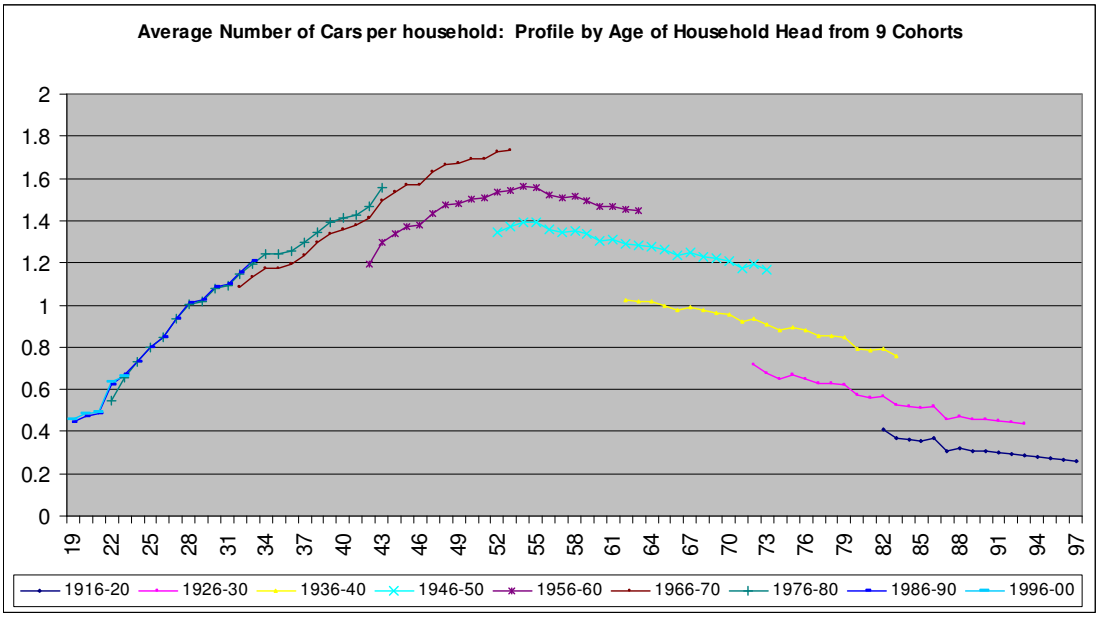
* National Transport Model

The forecasts in the early NRTF (1997) are the lowest, and all other studies predict higher car numbers in 2021. The early National Transport Model forecasts (Whelan,

2003) appear to be too high and have been subsequently revised down in Whelan (2007). In the current study, the forecasts based on nonlinear models are slightly lower than the latter but slightly higher than RAC (2002b). Overall, our nonlinear model results are comparable to the latest "official" figures. On the other hand, the forecasts based on our linear models are substantially higher than other studies (except Whelan, 2003), which seems to reinforce our concern that linear model will result in over-prediction as saturation can not be properly controlled.

Another "sense check" of the forecasts is to examine the profile of car ownership by cohort age. Similar to the profile of projected household income (Figure 8-2), the number of cars owned by household should follow a hump shape life cycle. The results from linear model L1 are presented in Figure 8-4, which shows the average number of cars per household for nine selected cohorts between age 19 and 97.

Figure 8-4 Model L1: Average Number of Cars per Household, X-axis by cohort age



The life cycle of car ownership is clearly illustrated in Figure 8-4, which also shows car ownership peaks when the household head is in his early 50s. Generally, the younger cohorts would have higher car ownership compared to older ones at the same age; however, this is no longer the case for the new cohorts. Such result is consistent with our assumptions that the fixed effects of all new cohorts are the same as that of the youngest cohort in the sample. Comparing Figure 8-4 to the observed car ownership

profile from FES (Figure 3-2), it seems our forecasts correctly reproduce the age effects and (diminishing) generation effects in the observed data.

On the other hand, in the restrictive fixed effect model of L2, the car ownership levels for the new cohorts are substantially higher. For cohort born between 1996 and 2000, the model predicts on average one car per household when the household head reaches the age of 19. This figure seems to be unrealistically high. The life cycle profile of car ownership from Model L2 is presented in Figure A-1 in Appendix 1. Also included in the Appendix are two tables on the average number of cars per household for all 20 cohorts in the forecasting period. Table A-2 refers to Model L1 and Table A-3 refers to Model L2.

For nonlinear models, we examine the proportions of household owning cars rather than the average number of cars. Figure 8-5 presents the results of Model D1, which refers to households with at least one car. There are some similarities between Figure 8-5 and Figure 8-4 of the linear model L1. One is the apparent hump shape life cycle; the other is the increase of car ownership for younger cohorts at a given age except for the new cohorts, which is due to the similar assumptions on fixed effects of these cohorts. On the other hand, the 'hump' for Model D1 is much flatter than L1, indicating a strong effect of saturation regarding household owing at least one car[51].

Similar profile for Model D2 is presented in Figure A-2 in Appendix 1. It shows similar age effects and generation effects of car ownership probability (proportions), although younger cohorts *always* have higher proportions than older ones at a given age (for both existing and new cohorts). Table A-4 and Table A-5 in the Appendix give the full results of the proportions of household owning 1+ car for Model D1 and D2 respectively.

Figure 8-6 presents the profile on the proportions of households owning two or more cars conditional on owning the first one. Compared to Model D1 of one plus car, the

[51] Another notable feature of the D1 profile is the bigger fluctuation of forecasts for households at young and old ages. This is due to the larger marginal effects of the household type split variables in D1. As discussed in the previous section, the projection of type split variables solely relies on the historical data and there are substantial noises for young and old cohorts due to small sample size.

Figure 8-5 Model D1: Proportion of Households Owning 1+ Car, X-axis by cohort age

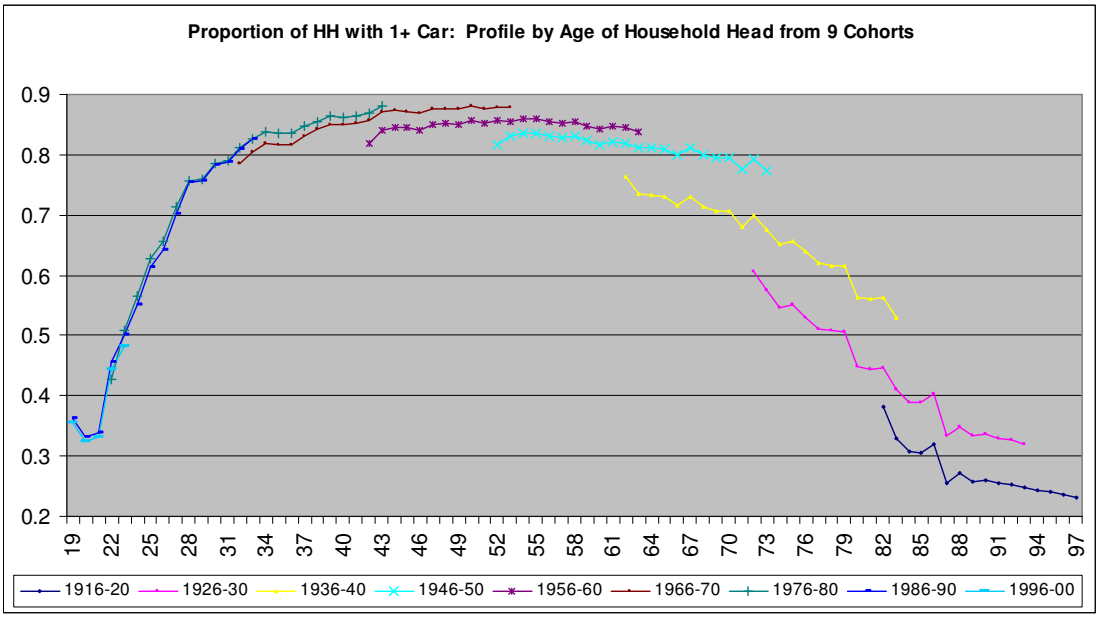
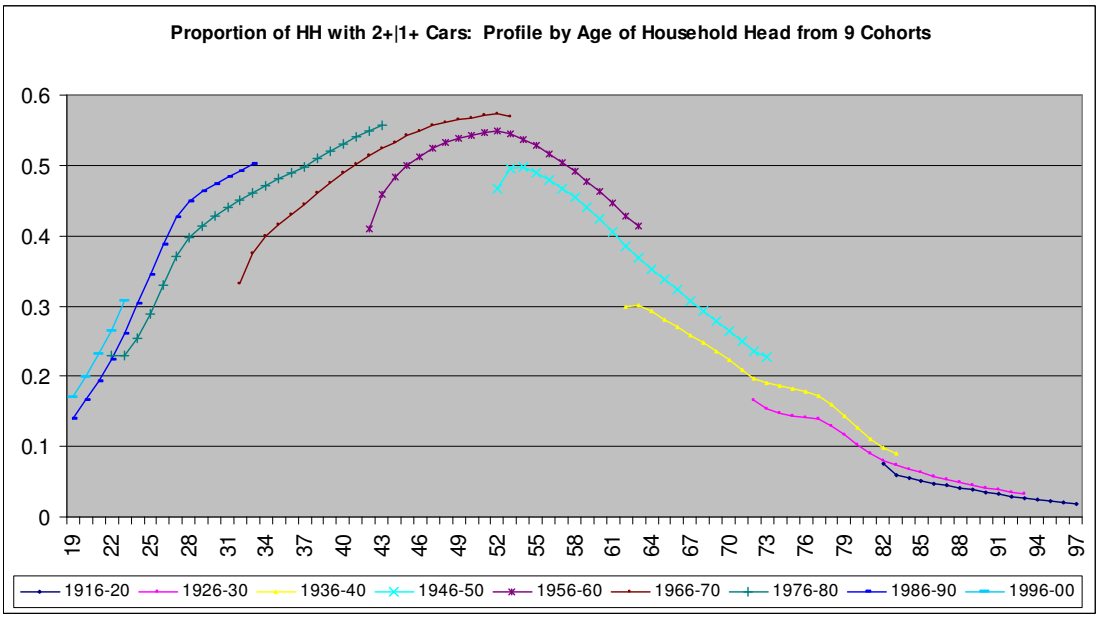


Figure 8-6 Model D3: Proportion of Households Owning 2+|1+ cars, X-axis by cohort age

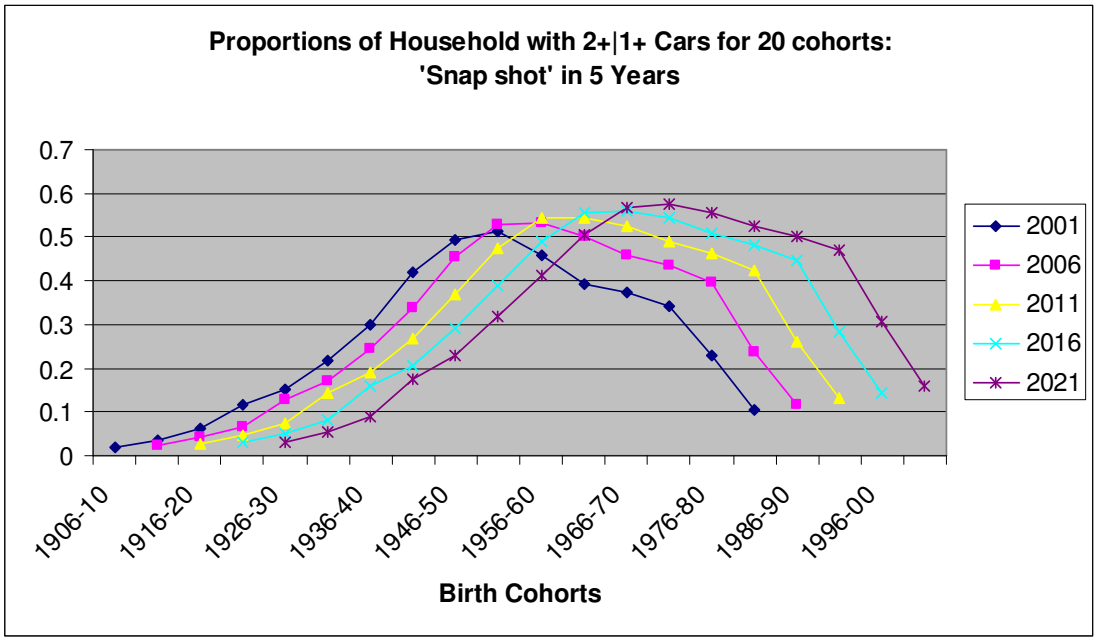


profile of D3 is more peaky since it is further away from saturation. Similar to Model D2, D3 is a pooled logit model so the differences of car ownership between younger and older cohorts at the same age are always present. The full results are also included in Table A-6 in the Appendix.

A better way to see saturation in force for model of Car 2+|1+ is to look at Figure 8-7. It contains data series for five cross sections of cohorts in five years, each reporting the

proportion of households with 2+|1+ cars. It illustrates the age effects and time trend (generation) effects from a different angle. One can also see the data series for the 2001 cross section is quite peaky, while those for the future years become flatter. This result indicates the gradual approach towards saturation over time.

Figure 8-7 Model D3: Proportion of Households with 2+|1+ cars, 5 cross sections of cohorts



A number of sensitivity tests have been carried out to ensure the forecasting models respond sensibly to the change of growth assumption and can be used to evaluate the impacts of transport policy measures. The first set of tests is income sensitivities, where we assume the GDP growth is 0.5% higher than the original assumption. The resulting forecasts as well as the implied elasticity from the five models are reported in Table 8-6, which should be compared to the central forecasts in Table 8-5. It shows that the two fixed effects models (L1 and D1) have lower income elasticity, mainly due to the smaller coefficient of the income variable in the econometric models. On the other hand, the pooled logit model D2 (combined with D3) has the highest income elasticity of 0.35. Overall, the income elasticity of these models appears to lie within a sensible range.

Other scenarios examined include no demographic change, real purchase price reduction of 0.87% per annum (instead of original assumption of 0.37%) and real running costs inflation of 0.5% per annum (rather than no costs change). The test

results for the nonlinear models are reported in Table 8-7, which should also be compared to the original results in Table 8-5.

Table 8-6 Sensitivity Test: GDP growth is 0.5% higher per annum

	L1	L2	D1+D3	D2+D3
2001	24973	24984	24975	24993
2006	28342	28882	27926	27962
2011	31430	32500	30314	30355
2016	34345	36031	32449	32531
2021	37169	39760	34635	34803
Elasticity	0.180	0.276	0.198	0.350

Table 8-7 Other Sensitivity Tests Results of Nonlinear Models

	D1+D3			D2+D3		
	Fixed Demography	Higher Price Drop	Run Costs Rise	Fixed Demography	Higher Price Drop	Run Costs Rise
2001	25113	24972	24943	25006	24967	24953
2006	27273	27898	27702	27122	27751	27642
2011	28735	30255	29905	28544	29961	29757
2016	29868	32358	31860	29653	31960	31662
2021	31019	34511	33874	30810	34069	33683
Elasticity	n.a.	-0.174	-0.062	n.a.	-0.128	*0.017*

The assumption of no demographic change leads to a big reduction of total car stock in 2021. Further examination reveals that most of the difference can be attributed to the rising number of households in the original forecasts. The higher deflation of purchase prices leads to a small increase of car numbers, which implies a purchase price elasticity of -0.17 and -0.13 for Model D1 and D2 (each combined with D3) respectively.

However, the results on running cost elasticity are unsatisfactory, as they are either very low or of wrong sign. This is mainly due to the fact that the coefficient of running costs variable has wrong sign in the Car 2+|1+ model, which reflects the concurrent rise of two plus car ownership and real car running costs in the 1990s. As a result, our model can not be used to evaluate the impacts of changing costs (and any transport policy measures that are designed to cause such change) on car ownership without fixing the running cost coefficient. This is the approach taken in the National Transport Model (Whelan, 2003; Whelan, 2007), where the parameters of both purchase price and running cost variables are constrained to certain values so that the model would

generate target elasticity. We have not implemented such measures since it is questionable of what 'target elasticity' should be.

8.3 Conclusion

In this chapter, selected econometric models estimated earlier have been applied to generate forecasts of car ownership in Great Britain to year 2021. A sub-model of input projection is developed to provide estimates of the household numbers and other explanatory variables in the forecasting period. A key feature of the input projection model is the ability to separate the age effects and time trend effects. For most of the explanatory variables, the time trend effects are reflected in the various growth rates applied to each of the 81 overlapping age band. The growth assumptions are obtained from various government sources and presented in a transparent way. Regarding the split of 8 household types within cohorts, the projections are solely based on the historical growth rate between different cohorts (at the same age) in the FES data. Overall, the projections of all input variables appear to be sensible.

Four sets of forecasts have been generated, two based on linear models and two based on (three) nonlinear models. To facilitate comparison, the model parameters have been slightly adjusted so that all 2001 forecasts are validated against the observed figures. The results are then compared to the observed total car stock in Great Britain between 2001 and 2006; beyond 2006, our forecasts are compared to other published sources. It shows our forecasts based on nonlinear models closely match the observed figures and are comparable to the latest "official" forecasts. On the other hand, both linear models forecast higher car ownership, presumably because the saturation effects are unaccounted for. In particular, if we ignore the "diminishing generation effects" and assume the cohort fixed effects are linear, the forecasted car stock becomes much higher in the distant future year.

A number of sensitivity tests have also been carried out. In general, the income elasticity in all four forecasting models appears to be sensible. While the purchase price elasticity also lies in the acceptable range, the running costs elasticity is either very low or with wrong sign for both nonlinear forecasting models. Although it is possible to force the models to have "right" elasticity by fixing the coefficient of the running costs

variable, we refrain from doing so in the current study as modeler's judgments on pre-defined elasticity will inevitably introduce bias in any examination of policy measures.

Chapter 9 Conclusion

Car ownership forecasting plays a central role in the planning and decision making of numerous public agencies and private organisations. It has been a lively area of research and numerous models have been constructed to forecast car demand. Traditionally the literature was dominated by static models, which rely on equilibrium assumptions that are sometimes questionable. Using dynamic models, on the other hand, it is possible to identify both the long run equilibrium conditions and short-term departure from such equilibrium. This in turn would reveal the true economic relationship and lead to more accurate forecasts.

The trend in car ownership modelling is to use dynamic and disaggregate methods. However, such effort has been hampered by the need for expensive and hard-to-collect panel data. To utilize the rich and readily available sources of long running cross sectional surveys, this study adopts the pseudo panel methods, which involves constructing an artificial panel based on (cohort) averages of repeated cross-sections. The cross-sectional data used here are the Family Expenditure Surveys between 1982 and 2000. By defining the cohorts on some time-invariant characteristics and developing appropriate econometric models, one could investigate dynamics for each cohort as well as heterogeneity between different cohorts. This would overcome the deficiencies in both the static models and aggregate time series.

The use of pseudo panel for car ownership modelling raises a number of interesting theoretical and empirical questions, which were initially listed in Chapter One. The results reported in this thesis provide satisfactory answers to these questions.

Chapter 4 and Chapter 5 deal exclusively with linear pseudo panel models. Firstly, it has been shown that the Weighted Least Square Estimator based on cohort means is equivalent to the Instrumental Variable (IV) estimator based on individual data from the micro survey and using cohort dummy as instruments. However, such relationship is based on the assumption that the economic relationship between the dependent variable and explanatory variables is linear and holds for individuals in the survey. Theoretically, this assumption is hard to defend for car ownership models, as individual household's car ownership decision is discrete. Empirically, this assumption is not

supported by the data, as the models using cohort averages of log-transformed variables have many coefficients that do not seem sensible, especially for the dynamic models. On the other hand, the models assuming linear economic relationship at cohort level produce much more satisfactory results.

We also investigate the frequently encountered problem of measurement errors in variables and the conditions required to ignore such problem. It has been shown that the way cohorts are constructed has direct implication on the bias of the within estimator if pseudo penal is to be estimated as genuine panel. The cohort should be defined in a way such that the population cohort means of the variables concerned vary as much as possible over time. Furthermore, the sample number in each cohort has to be sufficiently large to minimize sampling errors. These conditions appear to be met by our pseudo panel dataset constructed from the Family Expenditure Survey, which justify us to ignore the problem of measurement error in the empirical work.

Another important methodological issue that has been investigated in this thesis is the consistent estimation of dynamic models under different asymptotics. We first review the Error Corrected Within-Group Estimator, consistent when the time period is long ($T \to \infty$), and the Error Corrected GMM Estimator, consistent when the number of cohort is large ($C \to \infty$). We then present a Within-Group Estimator, which is computationally attractive and consistent under the most common asymptotic of $n_{ct} \to \infty$, which can be satisfied if the number of sample observations is sufficiently large for each cohort unit. Certain rank conditions have to be satisfied for identification, which require the cohort means of the dependent and independent variables should not exhibit perfect collinearity and vary over time. It is also required that there are at least three cross sections for the model to be identified.

The empirical models in this thesis incorporate various improvements to those in Dargay and Vythoulkas (1999) and the follow-on studies. Great efforts have been put in to identify the household structure (demographic characteristics) variables that best describe the data. The so-called life cycle effects are better captured by the second polynomial of the age of household head. We have investigated models with different functional forms, different representation of cohort effects and different assumptions of

error term. The robustness of the estimators has also been confirmed using parametric bootstrap technique. The final model has a very high adjusted R Square, showing a good degree of fit. All model coefficients have correct sign and sensible magnitudes and both the short run and long run elasticity of income and motoring costs lie in the range identified in the literature. If the dynamic model is viewed as a partial adjustment mechanism, the implied long run (equilibrium) effects are about 25% higher than the short run effects and full adjustment takes about four years. These results demonstrate key benefits of adopting the dynamic pseudo panel approach: to establish whether there is departure from equilibrium, the extent of departure and time taken towards full adjustment.

While the results of linear pseudo panel car ownership models are generally satisfactory, one important theoretical question remains. After rejecting the assumptions of linear economic relationship between car ownership level and various explanatory variables at individual household level, it is important to ask whether it is possible to develop a pseudo panel model that is consistent with the microeconomic theory of utility maximization. We provide a positive answer in this thesis.

Chapter 6 and 7 present a Random Utility Pseudo Panel Model, a theoretical model consistent with the Random Utility Theory. It combines the pseudo panel approach with discrete choice model, which does not seem to have been done before. We discuss the pros and cons of nonlinear (discrete choice) pseudo panel model and argue for its potential as an effective "third way" in modelling and forecasting using repeated cross section data. More specifically, it has the distinctive advantages of allowing both dynamics and saturation without the need for expensive genuine panel data. However, some valuable information on individual decision makers would be lost during cohort aggregation. On balance, it appears that nonlinear pseudo panel model is most suitable for forecasting purpose, while the case is less clear for analytical purpose.

Under the framework of random utility model (RUM), it is shown that the utility function of the pseudo panel model is a direct transformation from that of cross-sectional model and both share similar probability model albeit with different scale. In a standard random utility model of cross sectional data, the utility function consists of a deterministic term and a random term. For pseudo panel model, the deterministic term

can be further decomposed into three components including: sample mean observable utility, measurement error and individual decision maker's utility deviation from the cohort mean. We also assume the random part of the (standard) utility function has a "components of variance" structure, which is the sum of cohort specific component representing unobserved heterogeneity and a temporally independently identically distributed (IID) residual error component. Under the asymptotic of infinitive n_{ct} (the sample size per cohort is sufficiently large in each year), the measurement error converge in probability to zero. In the pseudo panel setting, the component that represents utility deviation from cohort mean has to be combined with the IID residual error term, which leads to models that has a similar probability function but with different scale compared to the cross-sectional model.

Based on this result, we then explore the various forms of true state dependence in the dynamic model and tackle the difficult econometric issues caused by the inclusion of lagged dependent variable. The fixed effect estimator is consistent only when the number of time period is sufficiently large, while the random effect estimator requires the orthogonality assumptions, i.e. the unobserved heterogeneity are uncorrelated with the explanatory variables. The mixed logit model allows all the parameters to be random, thus relaxing the orthogonality assumptions of the random effect model. The estimation of empirical models in the current study uses both the fixed effect and the mixed logit approaches, which have been implemented in a special Gauss routine for pseudo panels.

The empirical model of car ownership is formulated with a hierarchical structure. More specifically, two separate binary logit models are estimated. One predicts the probability of households owning at least one car (Car 1+), which uses the pseudo panel dataset constructed from the entire Family Expenditure Survey. The other predicts the probability of households owning two or more cars conditional on owning at least one car (Car 2+|1+), which is based on the second pseudo panel dataset constructed from a sub-sample of car owning households in FES. The use of hierarchical modeling structure avoids the undesirable Independence of Irrelevant Alternative (IIA) property of the multinomial logit model as well as the demanding computation tasks of the probit model.

This study specifically recognizes the importance of saturation in car ownership modelling. As a result, the discrete choice pseudo panel model has also been specified to model saturation. This leads to a so-called "Dogit" model, which is also consistent with the Random Utility Theory. The final models with the best fit for Car 1+ and Car 2+|1+ are both dynamic Dogit models. The estimated saturation levels are statistically significant and comparable with those identified in the literature. The short run and long run income elasticity appears to be sensible, with the elasticity for Car 2+|1+ much higher than that for Car 1+. It should be noted that the long run effects are indicative measures derived using linear Taylor approximation, which could be up to 70% higher than the short run effects[52]. This again demonstrates the benefits of allowing for dynamics in pseudo panel data.

Finally, this thesis presents the car ownership forecasts for Great Britain to year 2021. To the best of our knowledge, this is the first application of pseudo panel models to generate car ownership forecasts. Four sets of forecasting results are reported, two based on linear models and two based on discrete choice models. While the forecasts based on discrete choice models closely match the observed car stock between 2001 and 2006, those based on linear models appear to be too high. Furthermore, the results from nonlinear models are comparable to the findings in other authoritative studies, while the long term forecasts from linear models are significantly higher. These results highlight the importance of saturation, and hence the choice of model functional form, in car ownership forecasts.

In summary, this study has made significant contribution to the thriving research of car ownership by introducing the random utility pseudo panel model. It has been demonstrated that such model has solid theoretical foundation and strong empirical appeal. Further researches into its underlying economic meanings, statistical properties and estimation methods as well as more empirical applications in a wider context are likely to be most fruitful.

[52] For Pooled Logit Model of Car 1+, mid-income household. The differences between the long run and short run effects depend on various factors and could be as low as 10% for high income household in Model of Car 2+|1+,

Reference

Adda, J. and Cooper, R. (2000), Balladurette and Juppette: A Discrete Analysis of Scrapping Subsidies, *Journal of Political Economy*, 2000, vol. 108, (4), pp 778-806

Anderson, E. B. (1970), Asymptotic Properties of Conditional Maximum Likelihood Estimators, *Journal of Royal Statistical Society*, Series B, 32, pp283-301

Anderson, G. F. and Hussey, P. S. (2000), Population Aging: A Comparison among Industrialized Countries, *Health Affairs*, May/Jun 2000, Vol. 19 Issue 3, p191-204

Aptech Systems (1996), GAUSS Mathematical and Statistical System, Maple Valley, WA

Arellano, M. (2003), Discrete Choice with Panel Data, *Investigaciones Economicas*, Vol. XXVII (3), pp 423-458

Arellano, M. and Bond, S.R. (1991), Some Test of Specification for Panel Data: Monk Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, pp 277-297

Arellano, M. and Honore, B. (2001), Panel Data Models: Some Recent Developments *Handbook of Econometrics*, Vol. 5, (eds.) Heckman, J. and Learner, L. Elsevier Science

Arellano, M. and Carrasco, R. (2003), Binary Choice Panel Data Models with Predetermined Variables, *Journal of Econometrics*, 115, pp 125-157

Baldini, M. and Mazzaferro, C. (1999), *Demographic transition and Household Saving in Italy*, paper presented to the Bank of Italy Conference "Quantitative Research for Political Economy", Perugia, Dec. 1999

Baltagi, B.H. and Griffin, J.M. (1997), Pooled Estimators versus their Heterogeneous Counterparts in the context of dynamic demand for gasoline, *Journal of Econometrics*, 77, pp 303-327.

Baltagi, B. H., Bresson, G., Griffin J.M. and Pirotte, A. (2003), Homogeneous, heterogeneous or shrinkage estimators? Some empirical evidence from French regional gasoline consumption, *Empirical Economics*, 28, pp 795–811

Bates, J.J., Gunn, H., and Roberts, M., (1978), *A Disaggregate Model of Car Ownership*, DOE/DTp Research Report Number 20, HMSO, London

Beach, C. M. and Finnie, R. (2004), A Longitudinal Analysis of Earnings Change in Canada, *Canadian Journal of Economics*, Feb 2004, Vol. 37 Issue 1, pp219-241

Beck, N. (1991), Comparing Dynamic Specifications: The Case of Presidential Approval, *Political Analysis*, Vol. 3, pp 51–87

Beck, N., Epstein, D., Jackman, S. and O'Halloran, S. (2002), *Alternative Models of Dynamics in Binary Time-Series–Cross-Section Models: The Example of State Failure*, paper prepared for delivery at the 2001 Annual Meeting of the Society for Political Methodology, Emory University

Ben-Akiva, M.E. and Lerman, S.R. (1985), *Discrete Choice Analysis*, Cambridge, Massachusetts: MIT Press

Bhat, C. (1998), Accommodating Variations in Responsiveness to Level-of-Service Variables in Travel Mode Choice Models, *Transportation Research A* 32, pp455–507

Bhat, C. (2000), Incorporating Observed and Unobserved Heterogeneity in Urban Work Mode Choice Modeling, *Transportation Science* 34, pp228–238

Bhat, C.R. and Pulugurta, V. (1998), A Comparison of Two Alternative Behavioural Choice Mechanisms for Household Auto Ownership Decisions, *Transportation Research*, Part B, Volume 19B, pp 315-329.

Bierlaire, M. (1998), Discrete Choice Model, in M. Labbé, G. Laporte, K. Tanczos and Ph. Toint (eds), *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*, Vol. 166 of NATO ASI Series, Series F: Computer and Systems Sciences, Springer Verlag, pp. 203-227.

Biorn, E. (1992), The Bias of Some Estimator for Panel Data Models with Measurement Error, *Empirical Economics*, 17, pp 51-66

Boyd, H. and Mellman, R. (1980), The Effect of Fuel Economy Standards in the U.S. Auto Market: An Hedonic Demand Analysis", *Transportation Research*, Vol. 14A, No. 5-6, pp367-378

Bourguignon, F., Goh, C. and Kim, D. (2004), *Estimating individual vulnerability to poverty with pseudo-panel data*, World Bank Policy Research Working Paper 3375

Browning, M.. Deuton A. and Irish, M. (1985), A profitable approach to labour supply and commodity demand over the life-cycle, *Econometrica* 53, pp 503- 543

Brownstone, D. and Train, K. (1999), Forecasting New Porduct Penetration with Flexible Substitution Patterns, *Journal and Econometrics*, 89, pp109-129

Brownstone, D., Bunch, D. and Train, K. (2000) Joint Mixed Logit Models of Stated and Revealed Preferences for Alternative-Fuel Vehicles, *Transportation Research Part B: Methodological* Vol 34 No 5 pp315-338.

Butler, J. and Moffitt, R. (1982), A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model, *Econometrica*, 50, pp 761-764

Button, K.J., Pearman, A.D. and Fowkes, A.S (1982) *Car Ownership Modelling and Forecasting*, Gower, Aldershot, England

Button, K.J., Ngoe, N. and Hine, J. (1993) Modelling Vehicle Ownership and Use in Low Income Countries, *Journal of Transport Economics and Policy*, Jan. pp 51-67

Cardell, N.S. and Dunbar, F.C. (1980), "Measuring the Societal Impact of Automobile Downsizing", *Transportation Research,* Vol. 14A, No. 5-6, pp 423-434

Carro, J. (2003), *Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects*, CEMFI Working Paper No. 0304, CEMFI

Chamberlain, G. (1980), Analysis of Covariance with Qualitative Data, *Review of Economic Studies*, 47, pp 225-238

Chamberlain, G. (1984), Panel Data, in Griliches, Z. and Intriligator, M.D. eds. *Handbook of Econometrics*, 34, pp 305-334

Chamberlain, G. (1985), Heterogeneity, Omitted Variable Bias and Duration Dependence, in *Lognitudinal Analysis of Labor Market Data*, eds. Heckman, J. and Singer, B. No. 10 in Econometric Society Monographs Series, pp 3-38, Cambridge: Combridge University Press

Chen, S (1998), *Root-N consistent estimation of a panel data sample selection model*, unpublished manuscript, The Hong Kong University of Science and Technology.

Collado, M.D., 1997. Estimating Dynamic Models from Time Series of Independent Cross Sections, *Journal of Econometrics*, 82, pp 37-62

Cox, D. R. and Reid, N. (1987), Parameter Orthogonality and Approximate Conditional Inference, *Journal of Royal Statistical Society*, Series B, 49, pp 1-39

Cramer, J.S and Vos, A.(1985), *Een model voor prognoses van het personenautopark*, Interfaculteit der Actuariele wetenschappen en Econometrie, Universiteit van Amsterdam, mei 1985

Daganzo, C. F. (1979), *Multinomial Probit: The Theory and its Applications to Demand Forecasting*, New York: Academic Press

Daly, A. (1999), *How Much is Enough? Saturation Effects Using Choice Models*, Traffic Engineering and Control, Oct. 1999, pp 493-495

Daly, A.J. and Gunn, H.F. (1985), *Cost-Effective Methods for National-Level demand Forecastin*', IATBR Conference, Noordwijk

Daly, A.J. and Ortuzar, J. de D. (1990), Forecasting and Data Aggregation: Theory and Practice, *Traffic Engineering and Control*, Vol. 31, pp 632-643

Dargay, J. (1998), *Modelling Car Ownership in France and the UK: a Pseudo-panel Approach*, paper presented to Conference on the Economics and Institutions of Transport, Borlänge, Sweden, May 1998

Dargay, J. (2001), The Effect of Income on Car Ownership: Evidence of Asymmetry, *Transportation Research Part A*, Volume 35, pp807-821

Dargay, J. (2002), Determinants of car ownership in rural and urban areas: a pseudo-panel analysis, *Transportation Research Part E: Logistics and Transportation Review* Volume 38, Issue 5, September 2002, pp351-366

Dargay, J. and Vythoulkas, P. (1999), Estimation of a Dynamic Car Ownership Model, A Pseudo-Panel Approach, *Journal of Transport Economics and Policy*, Vol. 33, Part 3, Sept. 1999, pp 287-302

Dargay, J. and Gately, D. (1999), Income's Effect on Car and Vehicle Ownership, Worldwide: 1960-2015, *Transportation Research, Part A,* Vol. 33, pp101-138

Dargay, J. and Hivert, L. (2005), *The Dynamics of Car Ownership in EU Countries: A Comparison Based on the European Household Panel Survey*, paper presented to the European Transport Conference, Oct. 2005

Deaton, A. (1985), "Panel Data from Time Series of Cross Sections", *Journal of Econometrics*, 30, pp109-26

Deaton, A. (1992), *Understanding Consumption*, Oxford University Press

Department of Transport (1978) *Report of the Advisory Committee on Trunk Road Assessment*, HMSO, London

DETR (2000), *Transport 10 Year Plan 2000,* http://www.dft.gov.uk/about/strategy/whitepapers/previous/transporttenyearplan2000

Department for Transport (2004), *Transport Trends*, HMSO, London

Department for Transport (2005), *Vehicle Licensing Statistics*, 2005, http://www.dft.gov.uk/pgr/statistics/datatablespublications/vehicles/licensing/vehiclelicensingstatistics2005b

Department for Transport (2006a), *Vehicle Licensing Statistics Release*, 2006, http://www.dft.gov.uk/pgr/statistics/datatablespublications/vehicles/licensing/vehiclelicensingstatrelease2006

Department for Transport (2006b), *Transport Statistics Bulletin, Vehicle Excise Duty Evasion*, 2006, http://www.dft.gov.uk/pgr/statistics/datatablespublications/vehicles/excisedutyevasion/vehicleexcisedutyevasion2006

Department for Transport (2006c), *Transport Statistics Great Britain*, 2006, http://www.dft.gov.uk/pgr/statistics/datatablespublications/tsgb/2006edition/

De Jong, G.C. (1989a) *Some joint models of car ownership and car use,* Ph.D. thesis, Faculty of Economic Science and Econometrics, University of Amsterdam.

De Jong, G.C. (1989b) *Simulating car cost changes using an indirect utility model of car ownership and car use*, paper presented at PTRC SAM 1989, PTRC, Brighton.

De Jong, G.C. (1993), *Car Ownership Forecasts for France*, Internal Memorandum, Hague Consulting Group, The Netherlands

De Jong, G.C. (1996) A Disaggregate Model System of Vehicle Duration, Type Choice and Use, *Transportation Research - Part B*, Vol. 30, pp 263-276

De Jong, G.C. de and R. Kitamura (1992) *A review of household dynamic vehicle ownership models: holdings models versus transactions models*, Paper presented at 20th PTRC Summer annual meeting, Seminar E, London.

De Jong, G.C. and Pommer, J. (1996) *A Competing Risks Model of Vehicle Replacement, Disposal and Acquisition*, presented at the European Transport Conference, 1996

De Jong, G.C., Fox, J., Daly, A. Pieters, M. and Smit, R. (2004), "Comparison of Car Ownership Models", *Transport Reviews*, Vol. 24, No. 4, pp379-408

Devereux, P. (2003), *Small sample bias in grouping estimators: application to labor supply*, working paper, Department of Economics, UCLA

Dubin, J. and McFadden, D. (1984), An Econometric Analysis of Residential Electric Appliance Holdings and Consumption, *Econometrica*, Vol. 52, 2, pp 345-362

Gallez, (1994), "Identifying the Long Term Dynamics of Car Ownership: a Demographic Approach", *Transport Reviews,* Vol. 14, pp83-102

Garner, B.R., Godley, S. H. and Funk, R. R. (2002), Evaluating Admission Alternatives in an Outpatient Substance Abuse Treatment Program for Adolescents, *Evaluation & Program Planning*, Aug 2002, Vol. 25 Issue 3, p287-295

Gaudry, M. and Dagenais, M. (1979), The Dogit Model, *Transportation Research* Part B, 13 (2), pp105-112

Gilbert, C.C.S. (1992), A Duration Model of Automobile Ownership, *Transportation Research Part B*, Vol. 26, (2), pp 97-114

Girma, S. (2000), A Quasi Differencing Approach to Dynamic Modelling from a Time Series of Independent Cross-Sections, *Journal of Econometrics,* Vol. 98, pp 365-383

Glied, S. (2002), Youth Tobacco Control: Reconciling Theory and Empirical Evidence*, Journal of Health Economics*, Jan 2002, Vol. 21 Issue 1, p117-136

Golounov, V., Dellaert, B. and Timmermans, H. (2001), *A dynamic lifetime utility model of car purchase behavior using revealed preference consumer panel data*, paper submitted for presentation at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002

Golounov, V., Dellaert, B., Soest, A. and Timmermans, H. (2004), *Mixed Logit Heterogeneous Intertemporal Lifetime Utility Model of Car Leasing Behavior Using Conjoint Choice Data*, working paper, Tilburg University

Goodwin, P. (1997), Have Panel Surveys Told Us Anything New? in Golob, T. F., Kitamura, R. and Long, L. (eds), *Panels for Transportation Planning*, Boston

Goodwing, P., Kitamura, R. and Meurs, H. (1990), Some Principles of Dynamic Analysis of Travel Behavior, in Jones, P. (eds) *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*, Aldershot: Avebury blz, pp 56-73

Government Actuary Office (2003), *Population Projections*, http://www.gad.gov.uk/Population/2003/gb/wgb035y.xls

Greene, W. (1995), *Limdep Manual*, *Version 7*, Econometrics Software

Greene, W. (2001a), *Fixed and Random Effects in Nonlinear Models*, Working paper, Department of Economics, New York University

Greene, W. (2001b), *Estimating Econometric Models with Fixed Effects*, Working paper, Department of Economics, New York University

Greene, W. (2002), *The Bias of the Fixed Effects Estimator in Nonlinear Model*, Working paper, Department of Economics, New York University

Greene, W. (2003), *Econometric Analysis*, 5th Edition, New Jersey: Pearson Education, Inc.

Greene, W. (2004), *Interpreting Estimated Parameters and Measuring Individual Heterogeneity in Random Coefficient Models*, Working paper, Department of Economics, New York University

Guadagni, P. M. and J. Little (1983), A Logit Model of Brand Choice Calibrated on Scanner Data, *Marketing Science*, No. 2 (summer), pp. 203-238

Hanly, M. and Dargay, J. (2000), *Car Ownership in Great Britain – A Panel Data Analysis*, paper presented to European Transport Conference, PTRC, Cambridge

HCG (1993), *Een gedisaggregeerd model van bezitsduur, typekeuze, jaarkilometrage en brandstofverbruik van personenauto's*, HCG Report 319-1; The Hague: Hague Consulting Group

HCG (1995a), *Further development of a dynamic vehicle transactions model*, HCG Report 5037-1, The Hague: Hague Consulting Group

HCG (1995b), *Sensitivity runs with a dynamic vehicle transactions model*, HCG Report 5037-2, The Hague: Hague Consulting Group

HCG (2000), *Sydney Car Ownership Models*, HCG Report 9009-3B, The Hague: Hague Consulting Group

HCG and TØI (1990), *A Model System to Predict Fuel Use and Emissions from Private Travel in Norway from 1985 to 2025*, Hague Consulting Group, The Netherlands

Heckman, J. (1981a), Statistical Models for Discrete Panel Data, in *Structural Analysis of Discrete Data with Econometric Applications*, Manski, C. and McFadden, D. (eds.), Cambridge: MIT Press

Heckman, J. (1981b), The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time – Discrete Data Stochastic Process, in *Structural Analysis of Discrete Data with Econometric Applications*, Manski, C. and McFadden, D. (eds.), Cambridge: MIT Press

Hensher, D. and Wrigley, N. (1986), Statistical Modelling of Discrete Choices in Discrete Time with Panel Data, in *Behavioral Research for Transport Policy*, pp 97-116, Utrecht

Hensher, D., Bernard, P.O., Smith, N.C. and Wilthorpe, F.W. (1989), An Empirical Model of Household Automobile holdings, *Applied Economics*, Vol. 21, pp35-57

Hensher, D.A. and Mannering, F. (1994) Hazard Based Duration Models and their Application to Transport Analysis, *Transport Reviews*, Vol. 14, (1), pp 63-82

Hess, S. and Polak, J.W. (2005), *Accounting for random taste heterogeneity in airport-choice modelling*, paper presented to the 84th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2005; forthcoming in the Transportation Research Record

Hess, S., Bierlaire, M. & Polak, J.W. (2006), Capturing taste heterogeneity and correlation structure with Mixed GEV models, in Scarpa, R. and Alberini, A. (eds.), *Applications of Simulation Methods in Environmental and Resource Economics*, Springer Publisher, Dordrecht, The Netherlands, chapter 4, pp 55-76

Honore, B. E. (2002), Non-Linear Models with Panel Data, *Portuguese Economic Journal*, Vol. 1 (2), pp 163-179

Honore, B. E. and Kyriazidou, E. (2000), Panel Data Discrete Choice Models with Lagged Dependent Variables, *Econometrica* 68, pp 839-874

Hsiao, C. (1986), *Econometric Analysis of Panel Data*, Cambridge University Press

Institute for Social and Economic Research (2006), *Quality Profile: British Household Panel Survey, Version 2.0, Wave 1-13, 1991-2003*, http://www.iser.essex.ac.uk/ulsc/bhps/quality-profiles/BHPS-QP-01-03-06-v2.pdf

Jones, J. and Landwehr, J. (1988), Removing Heterogeneity Bias from Logit Model Estimation, *Marketing Science*, 7, 1, pp 41-59

Kitamura, R. (1990), Panel Analysis in Transportation Planning: An Overview, *Transportation Research A*, 24 (6), pp 401-417

Koyck, L. (1954), *Distributed Lags and Investment Analysis*, Amsterdam: North-Holland

Lauer, C., (2003), Family Background, Cohort and Education: A French–German Comparison Based on A Multivariate Ordered Probit Model of Educational Attainment, *Labour Economics*, April 2003, Vol. 10 Issue 2, p231-252

Leth-Petersen, S. and Bjorner, T. B. (2005), *A Dynamic Random Effects Multinomial Logit Model of Household Car Ownership*, AKF working paper, Denmark

Long, L. (1997), Panels for Transportation Planning: Theoretical Issues and Empirical Challenges, in Golob, T. F., Kitamura, R. and Long, L. (eds), *Panels for Transportation Planning*, Boston

Madre, J.L. (1990), "Long Term Forecasting of Car Ownership and Use", in Jones, P. (ed.) *Developments in Dynamic and Activity Based Approaches to Travel Analysis*, Aldershot: Gower Publishing, Oxford Studies in Transport

Madre, J. L., Bussiere, Y. and Armoogun, J. (1995), Demographic Dynamics of Mobility in Urban Areas: a Case Study of Paris and Grenoble, *Proceeding of the 7th World Conference on Transport Research, 16-21 July 1995, Sydney, Australia*, Elsevier Science Ltd.

Madsen, E. (2005), Estimating Cointegrating Relationships for Cross Sections, *Econometrics Journal*, 8, pp 380-405

Magnac, T. *State Dependence and Heterogeneity in Youth Unemployment Histories*, Working paper, INRA and CREST, Paris

Manski, C.F. (1987), Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data, *Econometrica*, 55, pp 357-362

Marshall, P. (1992), Estimating Time-Dependent Means in Dynamic Models for Cross-sections of Time Series, *Empirical Economics*, 17, pp25-33

McFadden, D. and Reid, F.A. (1975), Aggregate Travel Demand Forecasting From Disaggregate Behavioral Models, *Transportation Research Record*, 534, pp 24-37

McFadden, D. and Train, K. (1997) *Mixed MNL Models for Discrete Response*, Department of Economics, UC Berkeley

McFadden, D. and Train, K. (2000), Mixed MNL Models of Discrete Response, *Journal of Applied Econometrics*, 15, pp447-470

McKenzie, D. (2004), Asymptotic Theory for Heterogeneous Dynamic Pseudo Panels, *Journal of Econometrics*, 120, pp 235-262

Medlock, K.B. and Soligo, R. (2002), Car Ownership and Economic Development with Forecasts to the Year 2015, *Journal of Transport Economics and Policy,* Vol. 36, Part 2, May 2002, pp 163-188

Meurs, H., van Eijk, T. and Goodwin, P. (1990), Dynamic Estimation of Public Transport Elasticities, in Jones, P. (eds) *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*, Aldershot: Avebury blz, pp 371-383

Meurs, H., Haaijer, R., Smit, R. and Geurts, K. (2006), *DYNAMO: A New Dynamic Automobile Market Model for the Netherlands*, paper presented at the European Transport Conference, Oct. 2006

Moffitt, R. (1993), Identification and Estimation of Dynamic Models with Time Series of Repeated Cross Sections, *Journal of Econometrics* 59, pp 99-123

Mohammadian, A. and Miller, E. J. (2003), Dynamic modeling of household automobile transactions, *Transportation Research Record*, no. 1831, pp. 98-105

Mohammadian, A. and Rashidi, T. H. (2007), *Modeling Household Vehicle Transaction Behavior: A Competing Risk Duration Approach*, Transportation Research Board Annual Meeting, Paper No. 07-2014

Mundlak, Y. (1978), On the pooling of time series and cross section data, *Econometrica*, 46, pp 69-85

Newey, W.K. (1994), The asymptotic variance of semiparametric estimators, *Econometrica*, 62, pp 1349-1382.

Neyman, J. and Scott, E.L. (1948), Consistent Estimates Based on Partially Consistent Observations, *Econometrica*, 16, pp 1-32

Nickell, S. (198l), Biases in dynamic models with fixed effects, *Econometrica* 49, pp 1417-1426

Nobile, A., Bhat, C. and Pas, E. (1996), *A Random Effects Multinomial Probit Model of Car Ownership Choice*, research paper of Duke University and University of Massachusetts at Amherst.

NRTF (1989), *National Road Traffic Forecasts (Great Britain) 1989*, HMSO, London

NRTF (1997), *National Road Traffic Forecasts (Great Britain) 1997, Working Paper No. 1, Car Ownership: Modelling and Forecasting,* Department of the Environment, Transport and the Regions

Office of Deputy Prime Minister (1999), *Projections of households in England 2021*, http://www.odpm.gov.uk/stellent/groups/odpm_housing/documents/page/odpm_house_604206.hcsp

Office of National Statistics (2001), *National Travel Survey,* http://www.statistics.gov.uk/ssd/surveys/national_travel_survey.asp

Office of National Statistics (2002a), *Family Expenditure Survey – UK*, http://www.statistics.gov.uk/StatBase/Source.asp?vlnk=1385&More=Y

Office of National Statistics (2002b), *Population Trend*, 107, Spring 2002, HMSO, London

Office of National Statistics (2005), *Focus on Family*, http://www.statistics.gov.uk/focuson/families/

Office of National Statistics (2007), *Gross National Income Data* http://www.statistics.gov.uk/STATBASE/tsdataset.asp?vlnk=205

Ortuzar, J. and Willumsen, L. (2001), *Modelling Transport*, 3rd Edition, London: John Wiley & Sons, Ltd

Page, M., Whelan, G. and Daly, A. (2000), Modelling the Factors which Influence New Car Purchasing, Paper presented to *European Transport Conference 2000*, PTRC, Cambridge

Propper, C., Rees, H. and Green, K. (2001), The Demand for Private Medical Insurance in the UK: A Cohort Analysis, *The Economic Journal*, 111, May 2001, pp180-200

RAC (2002a), *Motoring towards 2050*, RAC Foundation, http://www.racfoundation.org/files/rac_foundation_2050.pdf

RAC (2002b), *Motoring towards 2050, Appendix 2: Technical Details of Forecasting Procedure*, RAC Foundation

Ramjerdi, F., Rand, L. and Saetermo, I-A (2000), *Models for Car Ownership, Transactions and Vehicle Type*, Research Report 187:2, Department of Technology and Society, Lund University, http://www.tft.lth.se/kfbkonf/5ramjerdiRandSetermo.PDF

Rand (2002), *Audit of Car Ownership Models*, Rand Europe Report 01192, Jan. 2002, Hague

Revelt, D. and Train, K. (1998), Mixed Logit with Repeated Choices, *Review of Economics and Statistics*, 80, pp647-657

Romilly, P., Song, H. and Liu, X. (1998), Modelling and Forecasting Car Ownership in Britain, A Cointegration and General to Specific Approach, *Jounral of Transport Economics and Policy*, Vol. 32, Part 2, pp165-185

Romilly, P., Song, H. and Liu, X. (2001), Car Ownership and Use in Britain: a Comparison of the Empirical Results of Alternative Cointegration Estimation Methods and Forecasts, *Applied Economics*, Vol. 33, pp1803-1818

Scottish Executive (2002), *Household Projections for Scotland: 2000-Based*, http://www.scotland.gov.uk/stats/bulletins/00179-00.asp

Tanner, J.C. (1958), *An Analysis of Increases in Motor Vehicles in Great Britain*, Research Note RN/1631, Road Research Laboratory, Harmondsworth

Tanner, J.C. (1978), Long Term Forecasting of Vehicle Ownership and Road Traffic (with discussion), *Journal of the Royal Statistical Society*, Series A, Volume 14, pp14-63

Train, K. (1986), *Qualitative Choice Analysis, Theory, Econometrics, and an Application to Automobile Demand*, Cambridge, MA: The MIT Press

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press

Treasury (2007), *HM Treasury Weekly Economic Indicator Data Bank*, 1st May 2007, http://www.hm-treasury.gov.uk/economic_data_and_tools/latest_economic_indicators/data_indic_index.cfm

Varian, H. (1992), *Microeconomic Analysis*, 3rd edition, W.W. Norton & Co.

Verbeek, M. and Nijman, T. (1992), Can cohort data be treated as genuine panel data? *Empirical Economics* 17, pp 9–23.

Verbeek, M. and Nijman, T. (1993), Minimum MSE estimation of a regression model with fixed effects from a series of cross sections, *Journal of Econometrics* 59, pp 125–136.

Verbeek, M. and Vella, F. (2005), Estimating Dynamic Models from Repeated Cross-Sections, *Journal of Econometrics*, 127, pp 83-102

Weir, G. (2003), Self-employment in the UK labour market, *Labour Market Trends*, Sept. 2003, Vol. 111 Issue 9, p441-452

Whelan, G. (2001), *Methodological Advances in Modelling and Forecasting Car Ownership in Great Britain*, paper presented to European Transport Conference, PTRC, Cambridge

Whelan, G. (2003), *Modelling Car Ownership in Great Britain*, unpublished PhD thesis, University of Leeds

Whelan, G. (2007), Modelling car ownership in Great Britain, *Transportation Research* Part A 41 pp 205–219

Whelan, G., Wardman, M. and Daly, A. (2000), *Is There a Limit to Car Ownership Growth? An Exploration of Household Saturation Levels Using two Novel Approaches*, paper presented to European Transport Conference, PTRC, Cambridge

Wooldridge, J. (2005), Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity, *Journal of Applied Econometrics*, Vol. 20, 1, pp 39-54

Yamamoto, T., Kitamura, R. and Kimura, S. (1999), Competing-Risks-Duration Model of Household Vehicle Transactions with Indicators of Changes in Explanatory Variables, *Transportation Research Record*, Volume 1676 / 1999, pp 116-123

Appendix 1 Supplementary Tables and Figures

Table A-1 Constructing pseudo panel by household head's date of birth (mean age for all cohorts)

Born	*1976-1980*	*1971-1975*	*1966-1970*	*1961-1965*	*1956-1960*	*1951-1955*	*1946-1950*	*1941-1945*	*1936-1940*	*1931-1935*	*1926-1930*	*1921-1925*	*1916-1920*	*1911-1915*	*1906-1910*	*1901-1905*
Cohort	**16**	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
1982				19	24	29	34	39	44	49	54	59	64	69	74	79
1983				20	25	30	35	40	45	50	55	60	65	70	75	80
1984				21	26	31	36	41	46	51	56	61	66	71	76	81
1985				22	27	32	37	42	47	52	57	62	67	72	77	82
1986				23	28	33	38	43	48	53	58	63	68	73	78	83
1987			19	24	29	34	39	44	49	54	59	64	69	74	79	
1988			20	25	30	35	40	45	50	55	60	65	70	75	80	
1989			21	26	31	36	41	46	51	56	61	66	71	76	81	
1990			22	27	32	37	42	47	52	57	62	67	72	77	82	
1991			23	28	33	38	43	48	53	58	63	68	73	78	83	
1992		19	24	29	34	39	44	49	54	59	64	69	74	79	84	
1993		20	25	30	35	40	45	50	55	60	65	70	75	80	85	
1994		21	26	31	36	41	46	51	56	61	66	71	76	81	86	
1995		22	27	32	37	42	47	52	57	62	67	72	77	82	87	
1996		23	28	33	38	43	48	53	58	63	68	73	78	83		
1997		24	29	34	39	44	49	54	59	64	69	74	79	84		
1998	20	25	30	35	40	45	50	55	60	65	70	75	80	85		
1999	21	26	31	36	41	46	51	56	61	66	71	76	81	86		
2000	22	27	32	37	42	47	52	57	62	67	72	77	82	87		

Table A-2 Forecasted Average Number of Cars per Household (Model L1)

Born	*2001-06*	*1996-00*	*1991-95*	*1986-90*	*1981-85*	*1976-80*	*1971-75*	*1966-70*	*1961-65*	*1956-60*	*1951-55*	*1946-50*	*1941-45*	*1936-40*	*1931-35*	*1926-30*	*1921-25*	*1916-20*	*1911-15*	*1906-10*
ID	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**
2000					0.28	0.54	0.95	1.09	1.13	1.20	1.38	1.34	1.23	1.03	0.88	0.71	0.53	0.41	0.25	0.13
2001					0.36	0.66	0.99	1.13	1.19	1.29	1.39	1.37	1.25	1.02	0.86	0.68	0.54	0.37	0.21	0.13
2002					0.44	0.73	1.00	1.18	1.23	1.34	1.40	1.39	1.24	1.01	0.85	0.65	0.53	0.36	0.20	0.13
2003					0.46	0.80	1.06	1.18	1.25	1.37	1.41	1.39	1.21	1.00	0.84	0.67	0.48	0.36	0.20	0.12
2004					0.48	0.85	1.07	1.19	1.27	1.38	1.42	1.36	1.22	0.98	0.81	0.65	0.47	0.37	0.19	0.12
2005				0.29	0.62	0.94	1.13	1.24	1.31	1.44	1.45	1.34	1.20	0.99	0.83	0.63	0.48	0.30	0.19	0.11
2006				0.37	0.67	1.00	1.18	1.29	1.39	1.47	1.46	1.35	1.19	0.97	0.80	0.63	0.44	0.32	0.19	
2007				0.45	0.73	1.02	1.22	1.34	1.43	1.48	1.48	1.33	1.19	0.96	0.78	0.62	0.43	0.31	0.18	
2008				0.47	0.80	1.08	1.22	1.36	1.46	1.50	1.47	1.30	1.17	0.95	0.79	0.57	0.43	0.31	0.17	
2009				0.48	0.85	1.09	1.24	1.38	1.47	1.51	1.44	1.31	1.15	0.92	0.77	0.56	0.43	0.30	0.17	
2010			0.29	0.62	0.94	1.15	1.28	1.41	1.52	1.54	1.42	1.29	1.16	0.93	0.75	0.57	0.37	0.30	0.16	
2011			0.38	0.67	1.01	1.20	1.33	1.50	1.56	1.54	1.43	1.28	1.14	0.91	0.75	0.53	0.39	0.29		
2012			0.45	0.73	1.02	1.24	1.38	1.54	1.57	1.56	1.41	1.28	1.13	0.88	0.74	0.52	0.37	0.28		
2013			0.48	0.80	1.08	1.24	1.40	1.57	1.59	1.56	1.38	1.26	1.12	0.90	0.69	0.51	0.37	0.28		
2014			0.49	0.85	1.09	1.25	1.41	1.57	1.59	1.53	1.38	1.24	1.08	0.88	0.68	0.52	0.36	0.27		
2015		0.29	0.63	0.93	1.15	1.30	1.45	1.63	1.62	1.51	1.36	1.25	1.10	0.86	0.69	0.46	0.36	0.26		
2016		0.38	0.66	1.01	1.20	1.35	1.54	1.66	1.63	1.52	1.36	1.23	1.07	0.85	0.65	0.47	0.35			
2017		0.46	0.73	1.03	1.24	1.39	1.58	1.67	1.65	1.49	1.35	1.22	1.05	0.84	0.64	0.46	0.35			
2018		0.48	0.80	1.09	1.24	1.41	1.61	1.69	1.65	1.46	1.34	1.21	1.06	0.80	0.63	0.46	0.34			
2019		0.49	0.84	1.10	1.26	1.43	1.61	1.69	1.61	1.47	1.32	1.18	1.04	0.79	0.64	0.45	0.33			
2020	0.33	0.63	0.93	1.15	1.30	1.47	1.67	1.73	1.59	1.45	1.33	1.19	1.02	0.79	0.58	0.45	0.33			
2021	0.42	0.66	1.01	1.21	1.35	1.56	1.70	1.73	1.60	1.45	1.31	1.16	1.02	0.76	0.60	0.44				

Note: Number in red is derived from Family Expenditure Survey;
Number in pink is estimated using parameters from the static car ownership model;

Table A-3 Forecasted Average Number of Cars per Household (Model L2)

Born	*2001-06*	*1996-00*	*1991-95*	*1986-90*	*1981-85*	*1976-80*	*1971-75*	*1966-70*	*1961-65*	*1956-60*	*1951-55*	*1946-50*	*1941-45*	*1936-40*	*1931-35*	*1926-30*	*1921-25*	*1916-20*	*1911-15*	*1906-10*
ID	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**
2000					0.51	0.54	0.95	1.09	1.13	1.20	1.38	1.34	1.23	1.03	0.88	0.71	0.53	0.41	0.25	0.08
2001					0.62	0.78	1.03	1.13	1.17	1.27	1.38	1.35	1.20	1.00	0.87	0.71	0.56	0.41	0.24	0.07
2002					0.70	0.90	1.08	1.16	1.21	1.32	1.40	1.34	1.18	0.99	0.85	0.70	0.56	0.40	0.23	0.05
2003					0.78	0.98	1.12	1.19	1.25	1.36	1.41	1.33	1.16	0.98	0.84	0.70	0.56	0.39	0.21	0.04
2004					0.84	1.05	1.15	1.23	1.29	1.40	1.43	1.32	1.14	0.97	0.83	0.69	0.55	0.37	0.20	0.02
2005				0.59	0.91	1.11	1.19	1.26	1.32	1.44	1.45	1.30	1.12	0.96	0.82	0.68	0.54	0.36	0.18	0.01
2006				0.72	0.97	1.16	1.23	1.30	1.37	1.47	1.45	1.28	1.12	0.96	0.81	0.68	0.52	0.34	0.16	
2007				0.81	1.04	1.20	1.26	1.34	1.41	1.49	1.43	1.27	1.11	0.95	0.80	0.67	0.51	0.33	0.15	
2008				0.88	1.10	1.23	1.29	1.37	1.44	1.50	1.42	1.25	1.10	0.93	0.80	0.65	0.49	0.31	0.13	
2009				0.95	1.16	1.27	1.33	1.41	1.48	1.52	1.40	1.23	1.09	0.92	0.79	0.64	0.47	0.29	0.11	
2010			0.67	1.01	1.22	1.30	1.36	1.45	1.52	1.53	1.38	1.21	1.08	0.91	0.78	0.63	0.46	0.27	0.09	
2011			0.81	1.07	1.26	1.33	1.39	1.48	1.54	1.53	1.36	1.20	1.07	0.90	0.77	0.61	0.44	0.25		
2012			0.90	1.14	1.30	1.37	1.43	1.52	1.55	1.51	1.34	1.19	1.06	0.89	0.75	0.59	0.42	0.24		
2013			0.98	1.20	1.33	1.40	1.46	1.56	1.57	1.49	1.32	1.18	1.05	0.88	0.74	0.58	0.40	0.22		
2014			1.04	1.26	1.37	1.43	1.50	1.59	1.58	1.47	1.30	1.17	1.03	0.87	0.73	0.56	0.38	0.20		
2015		0.75	1.10	1.31	1.40	1.46	1.53	1.62	1.59	1.45	1.28	1.16	1.02	0.86	0.71	0.54	0.36	0.19		
2016		0.90	1.17	1.36	1.43	1.49	1.57	1.64	1.59	1.43	1.27	1.15	1.01	0.85	0.70	0.52	0.35			
2017		1.00	1.23	1.39	1.46	1.53	1.60	1.66	1.57	1.42	1.26	1.14	1.00	0.84	0.68	0.50	0.33			
2018		1.07	1.29	1.43	1.50	1.56	1.64	1.67	1.55	1.40	1.26	1.13	0.99	0.82	0.66	0.49	0.31			
2019		1.14	1.35	1.46	1.53	1.60	1.67	1.68	1.53	1.38	1.25	1.12	0.98	0.81	0.65	0.47	0.30			
2020	0.83	1.20	1.41	1.50	1.56	1.63	1.70	1.70	1.51	1.36	1.24	1.10	0.98	0.80	0.63	0.45	0.28			
2021	0.99	1.26	1.45	1.53	1.59	1.66	1.72	1.69	1.50	1.35	1.23	1.09	0.97	0.78	0.61	0.43				

Note: Number in red is derived from Family Expenditure Survey;
Number in pink is estimated using parameters from the static car ownership model;

Table A-4 Forecasted proportion of household owning at least one car (Model D1)

Born	*2001-06*	*1996-00*	*1991-95*	*1986-90*	*1981-85*	*1976-80*	*1971-75*	*1966-70*	*1961-65*	*1956-60*	*1951-55*	*1946-50*	*1941-45*	*1936-40*	*1931-35*	*1926-30*	*1921-25*	*1916-20*	*1911-15*	*1906-10*
ID	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**
2000					0.33	0.43	0.71	0.79	0.82	0.82	0.82	0.82	0.80	0.76	0.71	0.61	0.46	0.38	0.25	0.16
2001					0.31	0.51	0.74	0.80	0.82	0.84	0.83	0.83	0.80	0.74	0.68	0.58	0.46	0.33	0.21	0.17
2002					0.36	0.57	0.73	0.82	0.83	0.84	0.83	0.84	0.79	0.73	0.67	0.55	0.45	0.31	0.20	0.17
2003					0.33	0.63	0.76	0.82	0.83	0.84	0.84	0.84	0.79	0.73	0.67	0.55	0.40	0.30	0.20	0.17
2004					0.34	0.66	0.77	0.82	0.84	0.84	0.84	0.83	0.79	0.72	0.64	0.53	0.39	0.32	0.19	0.16
2005				0.36	0.46	0.71	0.79	0.83	0.84	0.85	0.84	0.83	0.79	0.73	0.66	0.51	0.40	0.26	0.19	0.16
2006				0.31	0.51	0.76	0.81	0.84	0.86	0.85	0.84	0.83	0.78	0.71	0.64	0.51	0.36	0.27	0.19	
2007				0.36	0.56	0.76	0.82	0.85	0.86	0.85	0.84	0.82	0.78	0.71	0.61	0.51	0.34	0.26	0.19	
2008				0.33	0.62	0.79	0.82	0.85	0.86	0.86	0.84	0.82	0.78	0.71	0.62	0.45	0.34	0.26	0.18	
2009				0.34	0.65	0.79	0.82	0.85	0.86	0.85	0.84	0.82	0.77	0.68	0.60	0.44	0.36	0.25	0.18	
2010			0.37	0.46	0.71	0.81	0.84	0.86	0.86	0.86	0.84	0.82	0.78	0.70	0.58	0.45	0.29	0.25	0.18	
2011			0.32	0.50	0.76	0.83	0.85	0.87	0.86	0.85	0.84	0.81	0.76	0.68	0.57	0.41	0.30	0.25		
2012			0.36	0.55	0.76	0.84	0.85	0.87	0.86	0.86	0.83	0.81	0.76	0.65	0.57	0.39	0.29	0.24		
2013			0.33	0.61	0.78	0.83	0.85	0.87	0.87	0.86	0.83	0.81	0.76	0.66	0.52	0.39	0.29	0.24		
2014			0.33	0.64	0.79	0.83	0.86	0.87	0.87	0.85	0.83	0.80	0.74	0.64	0.51	0.40	0.28	0.24		
2015		0.39	0.45	0.70	0.81	0.85	0.86	0.88	0.87	0.85	0.83	0.81	0.75	0.62	0.52	0.33	0.28	0.23		
2016		0.32	0.49	0.75	0.83	0.86	0.87	0.88	0.87	0.85	0.82	0.80	0.73	0.62	0.48	0.35	0.28			
2017		0.36	0.54	0.76	0.84	0.86	0.88	0.87	0.87	0.85	0.82	0.79	0.71	0.62	0.46	0.33	0.27			
2018		0.33	0.60	0.78	0.83	0.86	0.88	0.88	0.87	0.84	0.82	0.80	0.72	0.56	0.46	0.34	0.27			
2019		0.33	0.63	0.79	0.83	0.87	0.87	0.88	0.87	0.85	0.81	0.78	0.70	0.56	0.47	0.33	0.26			
2020	0.43	0.44	0.69	0.81	0.85	0.87	0.88	0.88	0.86	0.85	0.82	0.79	0.69	0.56	0.40	0.33	0.26			
2021	0.37	0.48	0.75	0.83	0.86	0.88	0.88	0.88	0.87	0.84	0.81	0.77	0.68	0.53	0.42	0.32				

Note: Number in red is derived from Family Expenditure Survey;
Number in pink is estimated using parameters from the static car ownership model;

Table A-5 Forecasted proportion of household owning at least one car (Model D2)

Born	*2001-06*	*1996-00*	*1991-95*	*1986-90*	*1981-85*	*1976-80*	*1971-75*	*1966-70*	*1961-65*	*1956-60*	*1951-55*	*1946-50*	*1941-45*	*1936-40*	*1931-35*	*1926-30*	*1921-25*	*1916-20*	*1911-15*	*1906-10*
ID	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**
2000					0.47	0.43	0.71	0.79	0.82	0.82	0.82	0.82	0.80	0.76	0.71	0.61	0.46	0.38	0.25	0.41
2001					0.47	0.56	0.74	0.79	0.82	0.83	0.84	0.83	0.79	0.73	0.68	0.58	0.46	0.34	0.22	0.21
2002					0.49	0.64	0.76	0.80	0.82	0.84	0.84	0.82	0.78	0.72	0.67	0.57	0.44	0.31	0.20	0.14
2003					0.53	0.69	0.77	0.81	0.83	0.84	0.84	0.82	0.77	0.71	0.65	0.55	0.42	0.29	0.19	0.12
2004					0.58	0.73	0.79	0.82	0.83	0.85	0.84	0.81	0.76	0.71	0.64	0.54	0.41	0.27	0.17	0.11
2005				0.48	0.62	0.76	0.80	0.83	0.84	0.85	0.85	0.80	0.76	0.71	0.63	0.52	0.39	0.25	0.16	0.11
2006				0.50	0.66	0.77	0.81	0.84	0.84	0.85	0.84	0.80	0.75	0.70	0.62	0.51	0.37	0.24	0.15	
2007				0.53	0.70	0.79	0.82	0.84	0.85	0.85	0.84	0.79	0.75	0.69	0.60	0.49	0.35	0.22	0.14	
2008				0.58	0.73	0.80	0.83	0.85	0.85	0.85	0.83	0.79	0.75	0.68	0.59	0.47	0.32	0.21	0.13	
2009				0.62	0.76	0.81	0.83	0.85	0.85	0.85	0.82	0.78	0.75	0.67	0.58	0.45	0.30	0.19	0.12	
2010			0.48	0.66	0.78	0.82	0.84	0.85	0.86	0.85	0.81	0.77	0.74	0.65	0.56	0.42	0.28	0.18	0.12	
2011			0.52	0.69	0.79	0.82	0.85	0.86	0.86	0.85	0.81	0.77	0.74	0.64	0.54	0.40	0.26	0.17		
2012			0.57	0.73	0.80	0.83	0.85	0.86	0.86	0.84	0.80	0.77	0.73	0.63	0.52	0.37	0.24	0.16		
2013			0.61	0.76	0.81	0.84	0.86	0.86	0.86	0.84	0.80	0.76	0.72	0.61	0.50	0.35	0.23	0.15		
2014			0.65	0.78	0.82	0.85	0.86	0.86	0.86	0.83	0.79	0.76	0.70	0.60	0.48	0.33	0.21	0.14		
2015		0.49	0.69	0.80	0.83	0.85	0.86	0.87	0.86	0.82	0.78	0.76	0.69	0.59	0.45	0.31	0.20	0.13		
2016		0.54	0.72	0.81	0.84	0.86	0.86	0.87	0.85	0.82	0.78	0.75	0.68	0.57	0.43	0.28	0.18			
2017		0.60	0.75	0.82	0.84	0.86	0.87	0.87	0.85	0.81	0.78	0.74	0.67	0.55	0.40	0.27	0.17			
2018		0.64	0.78	0.83	0.85	0.87	0.87	0.87	0.84	0.81	0.78	0.73	0.66	0.53	0.38	0.25	0.16			
2019		0.68	0.80	0.84	0.86	0.87	0.87	0.87	0.83	0.80	0.77	0.72	0.65	0.51	0.36	0.23	0.15			
2020	0.50	0.72	0.81	0.84	0.86	0.87	0.87	0.87	0.83	0.79	0.77	0.71	0.63	0.48	0.33	0.21	0.14			
2021	0.57	0.75	0.82	0.85	0.87	0.87	0.87	0.86	0.82	0.79	0.76	0.70	0.62	0.46	0.31	0.20				

Note: Number in red is derived from Family Expenditure Survey;
Number in pink is estimated using parameters from the static car ownership model;

Table A-6 Forecasted proportion of household owning 2+|1+ cars (Model D3)

Born	*2001-06*	*1996-00*	*1991-95*	*1986-90*	*1981-85*	*1976-80*	*1971-75*	*1966-70*	*1961-65*	*1956-60*	*1951-55*	*1946-50*	*1941-45*	*1936-40*	*1931-35*	*1926-30*	*1921-25*	*1916-20*	*1911-15*	*1906-10*
ID	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**
2000					0.09	0.23	0.29	0.33	0.32	0.41	0.51	0.47	0.42	0.30	0.23	0.17	0.13	0.08	0.04	0.01
2001					0.10	0.23	0.34	0.38	0.39	0.46	0.51	0.49	0.42	0.30	0.22	0.15	0.12	0.06	0.04	0.02
2002					0.12	0.25	0.37	0.40	0.43	0.48	0.52	0.50	0.41	0.29	0.21	0.15	0.10	0.05	0.03	0.02
2003					0.15	0.29	0.39	0.42	0.45	0.50	0.52	0.49	0.39	0.28	0.19	0.14	0.09	0.05	0.03	0.02
2004					0.17	0.33	0.41	0.43	0.47	0.51	0.53	0.48	0.37	0.27	0.18	0.14	0.08	0.05	0.03	0.02
2005				0.10	0.20	0.37	0.42	0.44	0.49	0.52	0.53	0.47	0.35	0.26	0.17	0.14	0.07	0.04	0.03	0.02
2006				0.12	0.24	0.40	0.43	0.46	0.50	0.53	0.53	0.45	0.34	0.25	0.17	0.13	0.07	0.04	0.02	
2007				0.14	0.28	0.41	0.45	0.48	0.51	0.54	0.52	0.44	0.33	0.23	0.17	0.12	0.06	0.04	0.02	
2008				0.16	0.32	0.43	0.46	0.49	0.52	0.54	0.51	0.42	0.31	0.22	0.16	0.10	0.06	0.04	0.02	
2009				0.19	0.36	0.44	0.47	0.50	0.53	0.55	0.50	0.40	0.30	0.21	0.16	0.09	0.05	0.03	0.02	
2010			0.11	0.22	0.40	0.45	0.48	0.51	0.54	0.55	0.49	0.38	0.28	0.20	0.16	0.08	0.05	0.03	0.02	
2011			0.13	0.26	0.42	0.46	0.49	0.52	0.55	0.54	0.47	0.37	0.27	0.19	0.14	0.07	0.05	0.03		
2012			0.15	0.30	0.44	0.47	0.50	0.53	0.55	0.54	0.46	0.35	0.25	0.19	0.13	0.07	0.04	0.02		
2013			0.18	0.34	0.45	0.48	0.51	0.54	0.55	0.53	0.44	0.34	0.24	0.18	0.11	0.06	0.04	0.02		
2014			0.21	0.39	0.46	0.49	0.52	0.55	0.56	0.52	0.43	0.32	0.23	0.18	0.10	0.06	0.04	0.02		
2015		0.13	0.24	0.43	0.47	0.50	0.53	0.56	0.56	0.50	0.41	0.31	0.21	0.17	0.09	0.05	0.03	0.02		
2016		0.14	0.28	0.45	0.48	0.51	0.54	0.56	0.56	0.49	0.39	0.29	0.21	0.16	0.08	0.05	0.03			
2017		0.17	0.33	0.46	0.49	0.52	0.55	0.56	0.55	0.48	0.38	0.28	0.20	0.14	0.07	0.05	0.03			
2018		0.20	0.37	0.47	0.50	0.53	0.56	0.57	0.54	0.46	0.36	0.26	0.20	0.13	0.07	0.04	0.02			
2019		0.23	0.41	0.48	0.51	0.54	0.57	0.57	0.53	0.45	0.35	0.25	0.19	0.11	0.06	0.04	0.02			
2020	0.14	0.26	0.45	0.49	0.52	0.55	0.57	0.57	0.52	0.43	0.33	0.24	0.19	0.10	0.06	0.04	0.02			
2021	0.16	0.31	0.47	0.50	0.53	0.56	0.58	0.57	0.51	0.41	0.32	0.23	0.17	0.09	0.05	0.03				

Note: Number in red is derived from Family Expenditure Survey;
Number in pink is estimated using parameters from the static car ownership model;

Figure A-1 Model L2: Average Number of Cars per Household, X-axis by cohort age

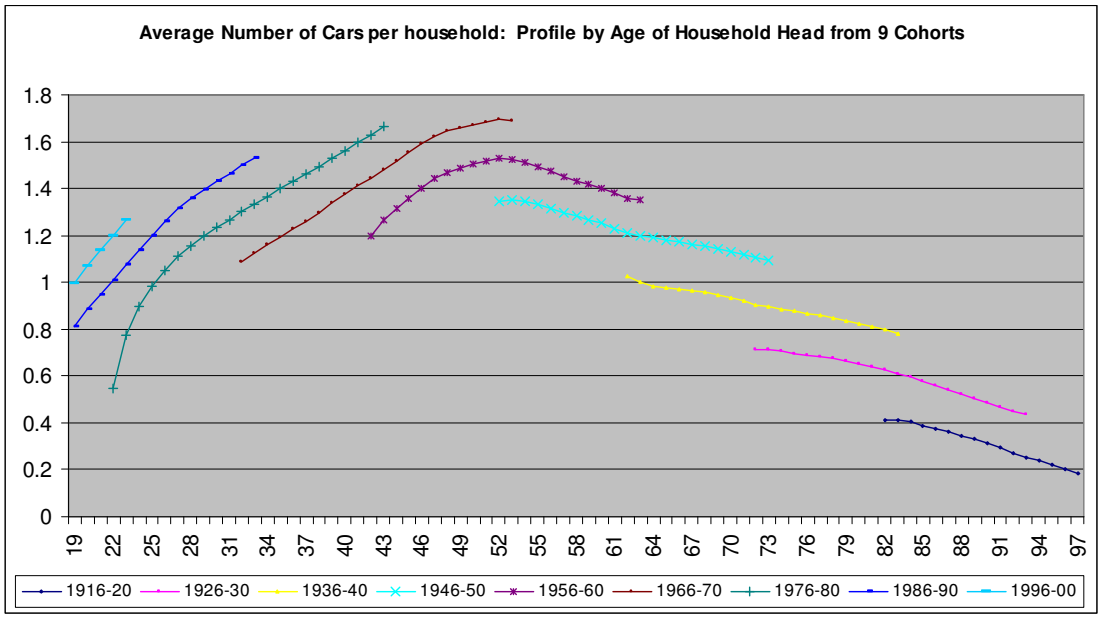
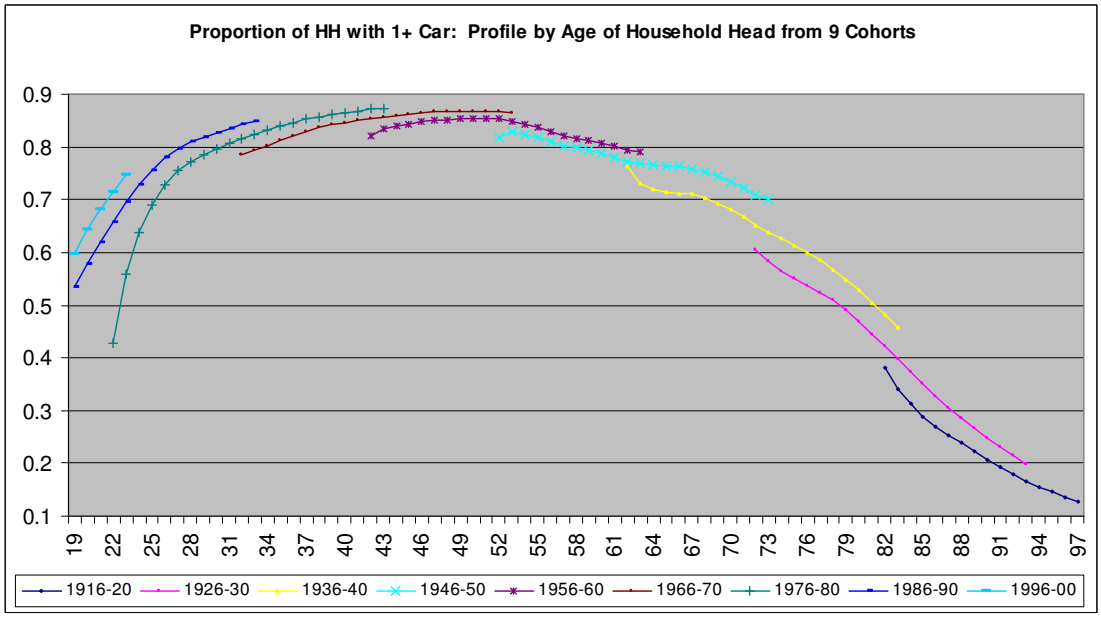


Figure A-2 Model D2: Proportion of Households with 1+ car, X-axis by cohort age



Appendix 2 Gauss Code for Pseudo Panel Mixed Logit Model

```
@Gauss code for Mixed Logit model of binary choice with proportions data@
@Code adapted from Revelt and Train, 1998@

new , 30000;
screen on;
output file=C:\temp\Gauss\C_choice8d_holton_out.txt reset;
print "This program estimates a mixed logit with fixed or normal coefficients";
nobs = 254;            @Number of choice situations.@
np=16;                 @Number of people (each facing multiple choice situations)@
nv=5;                          @Number of variables in the Xmat file@

load vars[nobs,nv]=C:\temp\Gauss\xmat2.dat;
load yvec[nobs,1]=C:\temp\Gauss\yvec.dat;
load times[np,1] =C:\temp\Gauss\times.dat;
load Wgt[nobs,1]=C:\temp\Gauss\weight.dat;

nrep=500;              @Number of draws to use in simulation.@

NFC=0;                 @Number of fixed variables@
IDFC={0};
NNC=5;                 @Number of normal variables@
IDNC={1,2,3,4,5};
@starting values in this order: mean for all fixed variables; mean1, std1, mean2, std2, etc. for all normal
variables@
b={-10.1146, 0, 1.8172, 0, 0.2108, 0, 0.0523, 0, -0.2044, 0};

    @ Create the Halton sequence @

    print "Creating Halton sequences ....";

        /* Number of random coefficients or 1, whichever is higher. */
        RowHlt = maxc( ones(1,1) | NNC );

     /* Provide prime number vector */

    prim = { 2 3 5 7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71
             73 79 83 89 97 101 103 107 109 113 };
    print "Halton sequences are based in primes: " prim[1,1:RowHlt];
    print;

   h = 1;
   hm = {};
   do while h<=RowHlt;
     hm1 = halton(10+nrep*np, prim[h]);
     hm1 = (cdfni(hm1))';
     @ The inverse-normal proc produces very extreme values sometimes. This truncates.@
     hm1=hm1.*(hm1 .le 10) + 10 .* (hm1 .gt 10);
     hm1=hm1.*(hm1 .ge -10) -10 .* (hm1 .lt -10);
     hm = hm|hm1[1,11:cols(hm1)];
     h=h+1;
   endo;

@Call maxlik, specify its globals, and do estimation.@
xx=ones(nobs,1);       @to use in maxlik@
library maxlik,pgraph;
#include maxlik.ext;
```

```
maxset;
_max_GradTol=0.000001;
_max_GradProc = &gr;
@_max_GradCheckTol=0.0001;@
_max_MaxIters=500;
_max_Algorithm=5;

print "Mixed Logit Panel estimation.";
{beta,f,g,cov,ret}=maxlik(xx,0,&llp,b);

call maxprt(beta,f,g,cov,ret);
format /m1 /rd 16,8;
print;
print "gradient(hessian-inverse)gradient is:" ((g/_max_FinalHess)'g);
print "diagonal of hessian:" ((diag(_max_FinalHess))');
print "Beta (8 decimal places)";
print beta;

/*Log-Likelihood routine for panel data.*/
proc llp(b,x);
@uses globals nobs,np,xmat,yvec,times,nrep@
local p,m,n,count,err,beta,ev,t,r,q1,q2,k,v,kmm;
p=zeros(np,1);          @Probability for np people's sequence of choices@

n=1;                  @To loop over people.@
count=0;               @Number of observations before n-th person.@

v=zeros(nobs,1);

 k = 1;
 do while k <= NFC;           @ Adds variables with fixed coefficients @
  v = v + b[k] * vars[.,IDFC[k,1]];
  k = k+1;
 endo;
do while n .le np;

  if NNC>0;
   err=hm[.,(nrep*(n-1)+1):(nrep*(n-1)+nrep)];
   r=zeros(1,nrep);                  @Hold probability for this person's sequence of choices (one for each
draw)@
   q1=ones(1,nrep);
   q2=ones(1,nrep);
  else;
   r=0;
   q1=1;
   q2=1;
  endif;

  t = 1;

  do while (t<=TIMES[n]);      @To loop over choice situations that this person faced.@
    kmm = count + t;
    ev = v[kmm,.];
    k = 1;
    do while k <= NNC;        @ Adds variables with normal coefficients @
      ev = ev + (b[NFC+2*k-1] + b[NFC+2*k] .* err[k,.])  @kth row, Nrep cols @
           .* Vars[kmm,IDNC[k,1]];
      k = k+1;
    endo;
```

```
      ev = exp(ev);
      @ev's for one choice situation: 1 by nrep@
      q1=yvec[kmm,1] .* Wgt[kmm,1] .* ln(ev[1,.] ./ (1 + ev[1,.]));
      q2=(1 - yvec[kmm,1]).* Wgt[kmm,1] .* ln(1 ./ (1 + ev[1,.]));
      r=r + q1 + q2;        @add log likelihood for previous choice situations: 1 by nrep@

      t=t+1;
    endo;
    p[n,1]=meanc(r');   @probability is average of r over all nrep draws@
    count=count+times[n];
    n=n+1;

endo;
retp(p);
endp;


/* GRADIENT PROCEDURE */
proc gr(b,x);

  @ Relies on the globals: NOBS, NP, NREP@
  @                        NFC, IDFC, NNC, IDFC @
  @                        Vars, Yvec  @

  local c, g, k, n, t, km, kmm, rd, nevar;
  local denom, der, v, ev, p0, p00, p1, err, mm;

@ Number of estimated variables @
nevar = NFC+NNC*2;

  v = zeros(NOBS,1);          @ Argument to logit formula      @
  p0  = zeros(NP,1);          @ Simulated probability          @
  der = zeros(NP,nevar);      @ Jacobian matrix                @

  k = 1;
  do while k <= NFC;          @ Adds variables with fixed coefficients @
    km = IDFC[k,1];
    v = v + b[k] .* Vars[.,km];
    k = k+1;
  endo;

  rd = 0;
  n = 1;
  do while n <= NP;

   if NNC>0;
    err=hm[.,(nrep*(n-1)+1):(nrep*(n-1)+nrep)]; @ err has nnc (or one) rows and NREP columns for each
person. @
    p00 = ones(1, NREP);
    p1  = ones(1, NREP);
    g = zeros(nevar, NREP);
   else;
    p00 = 1;
    p1  = 1;
    g = zeros(nevar, 1);
   endif;

    t=1;
    do while (t<=TIMES[n]);
      kmm = rd + t;
```

```
    ev = v[kmm,.];

    k = 1;
    do while k <= NNC;        @ Adds variables with normal @
      km = IDNC[k,1];
      ev = ev + (b[NFC+2*k-1] + b[NFC+2*k] .* err[k,.])
           .* Vars[kmm,km];
      k = k+1;
    endo;

    ev = exp(ev);
    denom = ev[1,.] + 1;
    p00   = p00 .* (ev[1,.] ./ denom);
    p1    = ev[1,.] ./ denom;

    k = 1;                  @Add (1-p)*x for fixed variables@
    do while k<=NFC;
      km = IDFC[k,1];
      g[k,.] = g[k,.] + Wgt[kmm,1] .* (yvec[kmm,1] - p1) .* Vars[kmm,km];
      k = k + 1;
    endo;

    k = 1;                  @Add (1-p)*x for normal variables@
    do while k<=NNC;
      km = IDNC[k,1];
      g[NFC+2*k-1,.] = g[NFC+2*k-1,.]
      + Wgt[kmm,1] .* (yvec[kmm,1] - p1) .* Vars[kmm,km];
      g[NFC+2*k,.]   = g[NFC+2*k,.]
      + Wgt[kmm,1] .* (yvec[kmm,1] - p1) .* Vars[kmm,km] .* err[k,.];
      k = k + 1;
    endo;

    t = t+1;
  endo;

  der[n,.] = (meanc((g)'))';
  rd = rd + TIMES[n];
  n = n + 1;
 endo;

 retp(der);

endp;

/* Halton procedure */

proc halton(n,s);
local phi,i,j,y,x,k;
k=floor(ln(n+1) ./ ln(s));    @We create n+1 Halton numbers including the initial zero.@
phi={0};
i=1;
do while i .le k;
 x=phi;
  j=1;
 do while j .lt s;
    y=phi+(j/s^i);
    x=x|y;
    j=j+1;
 endo;
 phi=x;
```

```
  i=i+1;
endo;

x=phi;
 j=1;
do while j .lt s .and rows(x) .lt (n+1);
   y=phi+(j/s^i);
   x=x|y;
   j=j+1;
 endo;

phi=x[2:(n+1),1];  @Starting at the second element gets rid of the initial zero.@
retp(phi);
endp;
```

Appendix 3 Deriving Long Run Elasticity Using Taylor Expansion

In a dynamic binary logit model, the probability of decision maker choosing Option 1 in year *t* can be expressed as:

$$P_t = \frac{\exp(\beta' x + \alpha P_{t-1})}{1 + \exp(\beta' x + \alpha P_{t-1})} \qquad (1)$$

Under long run equilibrium, the probability of owning car is stable over time, so the following equation holds:

$$P_t = P_{t-1} \qquad (2)$$

Substituting (1) into (2), it becomes: $P_{t-1} = \frac{1}{1 + \exp(-\beta' x - \alpha P_{t-1})}$ (3)

Dropping the sub-script $t-1$ and arrange terms in (3), we have:

$$(P^{-1} - 1)\exp(\alpha P) = \exp(-\beta' x) \qquad (4)$$

While it is not possible to directly solve for *P* based on (4), we start from using Taylor series to approximate the left hand side of (4), noting $f(P) = (P^{-1} - 1)\exp(\alpha P)$. The linear approximation of $f(P)$ is:

$$f(P) \approx [f(P_0) - f'(P_0)P_0] + f'(P_0) \cdot P \qquad (5)$$

where the subscript "0" indicates the function is evaluated at the arbitrarily chosen expansion point P_0. Also required is $f'(P_0)$, the first derivative of $f(P)$ evaluated at the expansion point:

$$f'(P_0) = \exp(\alpha P_0)(P_0^{-1}\alpha - P_0^{-2} - \alpha) \qquad (6)$$

Substituting (6) into (5) and collecting terms:

$$f(P) \approx \exp(\alpha P_0)[(P_0^{-1} - 1) - (P_0^{-1}\alpha - P_0^{-2} - \alpha)P_0 + (P_0^{-1}\alpha - P_0^{-2} - \alpha)P] \qquad (7)$$

Substituting (7) back into (left hand side) of (4) and dividing both sides by $\exp(\alpha P_0)$, we have:

$$(P_0^{-1} - 1) - (P_0^{-1}\alpha - P_0^{-2} - \alpha)P_0 + (P_0^{-1}\alpha - P_0^{-2} - \alpha)P \approx \exp(-\beta' x - \alpha P_0) \qquad (8)$$

Then one can easily solve (8) for:

$$P \approx \frac{\exp(-\beta' x - \alpha P_0) + (P_0^{-1}\alpha - P_0^{-2} - \alpha)P_0 - (P_0^{-1} - 1)}{P_0^{-1}\alpha - P_0^{-2} - \alpha} \tag{9}$$

Based on (9), the long run marginal effect of the explanatory variable x_k is:

$$\frac{\partial P}{\partial x_k} = \frac{1}{P_0^{-1}a - P_0^{-2} - a} \cdot \exp(-b'x - aP_0)(-b_k) \tag{10}$$

where b is the estimated coefficients for the exogenous explanatory variables and a is the estimated coefficient for the lagged dependent variable. The long run elasticity follows directly from the marginal effects.